RAS Techniques and Instruments

RASTAI **2,** 710–722 (2023) https://doi.org/10.1093/rasti/rzad051
Advance Access publication 2023 November 8

Bayesian model comparison for simulation-based inference

A. Spurio Mancini [1]★ M. M. Docherty,[1] M. A. Price[1] and J. D. McEwen[1,2]
[1]*Mullard Space Science Laboratory, University College London, Dorking RH5 6NT, UK*
[2]*Alan Turing Institute, London NW1 2DB, UK*

Accepted 2023 November 6. Received 2023 October 11; in original form 2022 July 15

ABSTRACT

Comparison of appropriate models to describe observational data is a fundamental task of science. The Bayesian model evidence, or marginal likelihood, is a computationally challenging, yet crucial, quantity to estimate to perform Bayesian model comparison. We introduce a methodology to compute the Bayesian model evidence in simulation-based inference (SBI) scenarios (often called likelihood-free inference). In particular, we leverage the recently proposed learned harmonic mean estimator and exploit the fact that it is decoupled from the method used to generate posterior samples, i.e. it requires posterior samples only, which may be generated by any approach. This flexibility, which is lacking in many alternative methods for computing the model evidence, allows us to develop SBI model comparison techniques for the three main neural density estimation approaches, including neural posterior estimation, neural likelihood estimation, and neural ratio estimation. We demonstrate and validate our SBI evidence calculation techniques on a range of inference problems, including a gravitational wave example. Moreover, we further validate the accuracy of the learned harmonic mean estimator, implemented in the HARMONIC software, in likelihood-based settings. These results highlight the potential of HARMONIC as a sampler-agnostic method to estimate the model evidence in both likelihood-based and simulation-based scenarios.

Key words: Machine Learning – Numerical Methods – Software – Statistics – Simulation-based Inference.

1. INTRODUCTION

Bayesian model comparison provides a robust and principled statistical framework for the selection of appropriate scientific models to describe observational data. The key quantity to perform model comparison in a Bayesian inference framework is the model evidence, or marginal likelihood, whose estimate allows one to assign relative weights to different models (see e.g. Trotta 2008 for a review of Bayesian model comparison, particularly in the context of cosmology). However, obtaining a precise and accurate estimate of the Bayesian model evidence is a computationally challenging task, involving a multidimensional integral which may quickly exceed the available computational resources for parameter spaces of even moderate dimensions. A variety of techniques for computing the Bayesian model evidence have been proposed (see e.g. Friel & Wyse 2012; Llorente et al. 2023 for reviews).

One of the most widely used classes of algorithms for estimating the model evidence, particularly in astrophysics and cosmology, is *nested sampling* (Skilling 2006), a method for which posterior inferences can also be computed as a byproduct (see e.g. Ashton et al. 2022; Buchner 2023 for reviews of nested sampling). Popular nested sampling algorithms, such as MULTINEST (Feroz & Hobson 2008; Feroz, Hobson & Bridges 2009) and POLYCHORD (Handley, Hobson & Lasenby 2015a, b) have been of remarkable success in multiple research areas. However, some of them can struggle in high-dimensional parameter spaces. The recently proposed *proximal nested sampling* framework scales to very high-dimensional settings (Cai, McEwen & Pereyra 2022). However, proximal nested sampling is restricted to log-convex likelihoods. Nevertheless, such likelihoods are common and so proximal nested sampling is likely to be particularly useful for inverse imaging problems. As the name suggests, nested sampling couples the computation of the model evidence to the sampling approach, restricting its flexibility.

The recently proposed *learned harmonic mean estimator* (McEwen et al. 2021) for computation of the model evidence removes this restriction. While the original harmonic mean estimator (Newton & Raftery 1994) can fail catastrophically since its variance may become very large and may not be finite (Neal 1994), the learned harmonic mean solves this problem by learning an approximation of the optimal internal importance sampling target distribution (McEwen et al. 2021). Critically, the learned harmonic mean estimator requires only samples from the posterior and so is agnostic to sampling strategy, affording it great flexibility, which is crucial to this work.

The need for efficient and reliable methods for computing the model evidence applies not only to likelihood-based settings, but also to *simulation-based inference* (SBI) frameworks. In the SBI setting (sometimes referred to as *likelihood-free inference*), the likelihood is either not available or too costly to be evaluated, and the inference process relies solely on the ability to simulate observables. *Approximate Bayesian computation* (ABC) is the traditional, prototypical SBI technique that relies on rejection sampling of parameter sets on the basis of a similarity metric between the simulated observables and the observations (see e.g. Beaumont 2019). However, ABC methods can easily require an unfeasibly large number of simulations to reach convergence, limiting their applicability. More recently *neural density estimation* techniques have been introduced to surrogate

★ E-mail: a.spuriomancini@ucl.ac.uk

densities directly. Such novel frameworks have seen numerous successful applications in various scientific areas, carrying great promise for the future due to their ability to avoid the evaluation of an explicit (and possibly incorrect) likelihood function, while limiting the number of simulations with clever use of cutting-edge machine learning algorithms. For a recent review of SBI techniques we refer the reader to Cranmer, Brehmer & Louppe (2020). Neural density estimation methods have recently found remarkable success in cosmology, with general-purpose open-source software readily available (e.g. PYDELFI[1] by Alsing et al. 2019 and SWYFT[2] by Cole et al. 2022).

The development of new SBI techniques has so far mostly concentrated on optimizing the task of parameter estimation, while model selection has received less attention. Nevertheless, model selection is a critical component of a complete statistical analysis, particularly in scientific fields where selection of the appropriate model is often the fundamental question. While there has been some consideration of model selection for SBI, the field remains nascent. Brewer, Pártay & Csányi (2011) propose a technique based on diffusive nested sampling to compute the model evidence for ABC. However, this approach is restricted to ABC, which as discussed above can be inefficient and of limited applicability, and is not straightforwardly generalizable to modern neural density estimation approaches.

In this article we introduce a methodology to compute the model evidence, in order to facilitate Bayesian model comparison, for modern neural density estimation approaches to SBI. Our methodology leverages the learned harmonic mean estimator (McEwen et al. 2021). We exploit the fact that the learned harmonic mean estimator is agnostic to sampling strategy and only requires samples from the posterior distribution. In some neural density estimation approaches samples can be generated directly (e.g. by pushing samples from a simple base distribution, such as a Gaussian, forward through a normalizing flow, Papamakarios et al. 2021). To support such neural density estimation approaches it is therefore essential that model evidence computation is decoupled from sampling strategy, as is the case with the learned harmonic mean estimator.

The remainder of this article is structured as follows. In Sections 2 and 3, we concisely review, respectively, Bayesian model comparison and neural density estimation approaches to SBI. In Section 4, we introduce our methodology to perform Bayesian model comparison in the context of SBI, leveraging the learned harmonic mean estimator (McEwen et al. 2021). We present algorithms to compute the Bayesian model evidence for each of the three main neural density estimation approaches to SBI that are reviewed by Cranmer et al. (2020). In Section 5, we report the results from numerical experiments that demonstrate and validate our methodology. Finally, we conclude in Section 6 with a review of our main findings.

2. BAYESIAN MODEL COMPARISON

We review here the fundamentals of Bayesian model comparison, focusing on the challenges associated with estimation of the model evidence. For a more extensive review we refer the reader to, e.g. Trotta (2008). We also summarize the key concepts underlying the learned harmonic mean estimator since it is an integral component of this work (we refer the reader to McEwen et al. 2021 for further details).

2.1. Bayesian model evidence

The definition of model evidence in a Bayesian statistical framework follows directly from Bayes' theorem. For a given model $\mathcal{M}$, Bayes' theorem describes the connection between the conditional probabilities of model parameters $\boldsymbol{\theta}$ and data $\boldsymbol{d}$:

$$p(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{M}) = \frac{p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\boldsymbol{d}|\mathcal{M})}, \tag{1}$$

where $p(\boldsymbol{\theta}|\boldsymbol{d}, \mathcal{M})$ is the posterior distribution of the parameters, given the observed data $\boldsymbol{d}$ and the assumed model $\mathcal{M}$, $p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M})$ is the likelihood function of the data $\boldsymbol{d}$ given parameters $\boldsymbol{\theta}$ and model $\mathcal{M}$, $p(\boldsymbol{\theta}|\mathcal{M})$ is the prior distribution of model parameters $\boldsymbol{\theta}$ for a given model $\mathcal{M}$, and $p(\boldsymbol{d}|\mathcal{M})$ is the model evidence, i.e. the probability of data $\boldsymbol{d}$ for a given model $\mathcal{M}$. The Bayesian model evidence is given by the normalization factor of the posterior distribution $p(\boldsymbol{\theta}|\mathcal{M}, \boldsymbol{d})$:

$$z = p(\boldsymbol{d}|\mathcal{M}) = \int \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M}). \tag{2}$$

Since the model evidence is a normalization factor for the posterior distribution and is independent of the model parameters, the evidence is usually disregarded in parameter estimation tasks. However, for model selection the evidence becomes the crucial quantity to compute. Being the integral of the likelihood over the prior (cf. equation 2), the evidence allows one to assign relative weights to different models. The evidence ratio between two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ enters the expression for the comparison of their posterior distributions, which again follows from Bayes' theorem:

$$\frac{p(\mathcal{M}_1|\boldsymbol{d})}{p(\mathcal{M}_2|\boldsymbol{d})} = \frac{p(\boldsymbol{d}|\mathcal{M}_1)p(\mathcal{M}_1)}{p(\boldsymbol{d}|\mathcal{M}_2)p(\mathcal{M}_2)}. \tag{3}$$

In many cases a priori probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$ of the two models are considered to be equal, hence the ratio of posterior distributions becomes equivalent to the evidence ratio or Bayes factor

$$B_{12} = \frac{p(\boldsymbol{d}|\mathcal{M}_1)}{p(\boldsymbol{d}|\mathcal{M}_2)} = \frac{z_1}{z_2}. \tag{4}$$

For notational brevity, henceforth we drop the explicit conditioning on models unless there are multiple models under consideration.

2.2. Algorithms for evidence estimation

Computing the evidence for a given model is numerically challenging due to the multidimensional integral in equation (2). Many techniques have been proposed to tackle this challenge, such as thermodynamic integration (e.g. Beltrán et al. 2005; Gregory 2005; Bridges, Lasenby & Hobson 2006), the Savage–Dickey density ratio (e.g. Trotta 2007), methods based on kth nearest-neighbour distances in parameter space (Heavens et al. 2017), nested sampling (Skilling 2006), and others (see e.g. Friel & Wyse 2012; Llorente et al. 2023).

Nested sampling reduces the computation of the evidence to the evaluation of a one-dimensional integral, and as a byproduct provides samples that can be used to compute posterior inferences, thus supporting both parameter estimation and model selection. Multimodal nested sampling, implemented in the MULTINEST[3] software (Feroz & Hobson 2008; Feroz et al. 2009; see also Buchner et al. 2014 for the Python wrapper PYMULTINEST[4]), has seen enormous success, with widespread application across multiple research fields,

[1] https://github.com/justinalsing/pydelfi
[2] https://github.com/undark-lab/swyft
[3] https://github.com/farhanferoz/MultiNest
[4] https://github.com/JohannesBuchner/PyMultiNest

as has the slice sampling nested sampling algorithm implemented in the POLYCHORD[5] software (Handley et al. 2015a, b). Proximal nested sampling (Cai et al. 2022) is implemented in the PROXNEST[6] software, which has only been released very recently but is likely to be particularly useful for inverse imaging problems.

Measuring the model evidence is a numerical process that, if repeated multiple times, produces a distribution of values. Ideally, these distributions would be narrow, and even more importantly they should provide unbiased estimates of the model evidence. However, this might not always be the case (Lemos et al. 2023). It is, therefore, crucial to develop alternative ways to estimate the model evidence, so as to perform cross-checks on the final model selection statements. Throughout we use the terminology of an *accurate* estimator as one with a low bias and a *precise* estimator as one with a low variance.

2.3. Learned harmonic mean estimator

The *original harmonic mean estimator* for computation of the Bayesian model evidence was first introduced by Newton & Raftery (1994). From Bayes' theorem it follows that the reciprocal evidence is related to the harmonic mean of the likelihood by

$$\rho \equiv \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{d})}\left[\frac{1}{p(\boldsymbol{d}|\boldsymbol{\theta})}\right] = \int \mathrm{d}\boldsymbol{\theta}\, \frac{1}{p(\boldsymbol{d}|\boldsymbol{\theta})}\, p(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{1}{z}. \tag{5}$$

This relation can be used to construct an estimator of the reciprocal evidence

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{p(\boldsymbol{d}|\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}|\boldsymbol{d}), \tag{6}$$

using N samples $\{\boldsymbol{\theta}_i\}_{i=1}^N$ of the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$. However, this estimator may present very large or even diverging variance (Neal 1994).

Gelfand & Dey (1994) proposed a modification to the original harmonic mean estimator, what we call the *retargeted harmonic mean estimator*, introducing a normalized target distribution $\varphi(\boldsymbol{\theta})$ to define the modified estimator

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\theta}_i)}{p(\boldsymbol{d}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}|\boldsymbol{d}), \tag{7}$$

from which the original harmonic mean estimator is recovered for $\varphi(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.

The original harmonic mean estimator can be interpreted as importance sampling, where the importance sampling distribution is the posterior and the target distribution is the prior. It is therefore not surprising that the original estimator suffers poor variance properties since the prior is typically broader than the posterior, whereas importance sampling requires the sampling density to be broader than the target. By introducing a new target $\varphi(\boldsymbol{\theta})$ this issue can be circumvented provided $\varphi(\boldsymbol{\theta})$ is narrower than the posterior. Critically, however, the introduced target must be a normalized probability distribution.

The question remains: how does one set the target distribution? A variety of approaches have been considered previously, however none have proved completely satisfactory. One approach is to consider a multivariate Gaussian (Chib 1995); however, such a target typically has tails that are too broad. An alternative is to consider indicator functions (Robert & Wraith 2009; van Haasteren 2014); however, for complicated posterior distributions these typically capture a small region of parameter space only and so are inefficient.

It was recognized by McEwen et al. (2021) that the optimal target distribution is the normalized posterior

$$\varphi^{\text{optimal}}(\boldsymbol{\theta}) = \frac{p(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{z}. \tag{8}$$

While this exact quantity is a priori inaccessible since it involves knowledge of the evidence z – precisely the quantity we are attempting to estimate – an approximation of $\varphi(\boldsymbol{\theta})$ can be derived from posterior samples by machine learning techniques. This is the rationale of the *learned* harmonic mean estimator of McEwen et al. (2021). Moreover, the learned approximation of the posterior need not be highly accurate; but critically it must have narrower tails than the posterior. Strategies to learn appropriate targets with these properties are presented in McEwen et al. (2021). In a nutshell, when learning models for $\varphi(\boldsymbol{\theta})$ a bespoke optimization problem is considered that penalizes the variance of the resulting learned harmonic mean estimator, while ensuring it is unbiased, with possible additional regularization. This effectively ensures the tails of $\varphi(\boldsymbol{\theta})$ are contained within the posterior. Thus, instead of learning a general distribution that matches the posterior, a distribution that is effective for the subsequent evidence computation is learned. For further details see McEwen et al. (2021). Note that $\varphi(\boldsymbol{\theta})$ can be trained simply from samples of the posterior and evaluating the posterior density is not strictly necessary. However, it is necessary to evaluate the normalized density of $\varphi(\boldsymbol{\theta})$ once it is trained. The learned harmonic mean estimator is implemented in the HARMONIC[7] software.

We conclude this section by highlighting that the learned harmonic mean estimator produces estimates of the evidence purely from samples of the posterior distribution; there is no requirement on the specific method used for sampling, i.e. HARMONIC is agnostic to the method used to generate posterior samples. As we shall see later in this work, this is the key property of the learned harmonic mean estimator that allows it to be used in a variety of SBI scenarios.

3. SIMULATION-BASED INFERENCE

We provide a brief overview of the main algorithms used for SBI, referring the reader to Cranmer et al. (2020) for a more extensive review. The focus of recent SBI developments and existing literature is on parameter estimation, hence we discuss SBI in this context. In Section 4, we introduce methodologies to perform Bayesian model comparison in an SBI setting.

The original SBI methodology, based on ABC (see e.g. Beaumont 2019), involves simulating realizations of the observables at each of the explored points in parameter space, and accepting or rejecting these points based on their similarity with the observed data, within a tolerance ϵ. This rejection sampling method recovers an accurate representation of the underlying density distribution in the limit $\epsilon \to 0$, at which point the low simulation efficiency makes computational costs infeasibly high for inference of parameter spaces with even moderate dimensionality. More recently, neural density estimation techniques have been introduced to overcome this computational limitation by increasing simulation efficiency. We focus the remainder of this article on neural density estimation approaches for SBI.

In contrast to ABC, neural density estimation leverages deep neural networks to approximate conditional probability densities and is able to speed up inference by orders of magnitude (Papamakarios &

[5] https://github.com/PolyChord/PolyChordLite
[6] https://github.com/astro-informatics/proxnest
[7] https://github.com/astro-informatics/harmonic

Murray 2016). Neural density estimation involves learning a conditional density estimator $q_{\boldsymbol{\phi}}$, parametrized by weights $\boldsymbol{\phi}$, to approximate a target distribution (either the posterior distribution, the likelihood function or a density ratio proportional to the likelihood) from a training set of N pairs of (typically) prior samples and simulations $\{\boldsymbol{\theta}_i, \boldsymbol{d}_i\}_{1=1}^{N}$. Provided the density estimator is sufficiently expressive, $q_{\boldsymbol{\phi}}$ will recover an accurate estimate of the target distribution in the limit $N \to \infty$.

Neural density estimation workflows have three main phases: simulation, training, and inference. The simulation phase generates the training pairs $\{\boldsymbol{\theta}_i, \boldsymbol{d}_i\}$ that are used in the training phase to tune the weights of the neural network $\boldsymbol{\phi}$ such that $q_{\boldsymbol{\phi}}$ approximates the target density. In the inference phase, $q_{\boldsymbol{\phi}}$ is then conditioned on a specific observation $\boldsymbol{d}_0$ and parameter inference is performed.

Single runs of the simulation and training phases amortizes the training of the density estimator, allowing offline inference to be run on multiple different observations, aptly named *amortized neural density estimation*. However, we are often interested in inference on a specific observation $\boldsymbol{d}_0$. For this, amortized neural density estimation tends to be inefficient as generating training pairs for the density estimator across the entire prior parameter support includes many points in parameter space with very low posterior density $p(\boldsymbol{\theta}|\boldsymbol{d}_0)$.

To rectify this simulation inefficiency one can run *sequential neural density estimation*, where multiple rounds of simulation and training are run sequentially to ensure there is a greater focus on regions of high posterior density. This is done by generating simulations from an alternative prior proposal distribution $\tilde{p}(\boldsymbol{\theta})$. This proposal distribution is iteratively updated between rounds such that for R rounds, the proposal posterior of the ith round $\tilde{p}_i(\boldsymbol{\theta}|\boldsymbol{d}_0)$ becomes the proposal prior for the subsequent round $\tilde{p}_{i+1}(\boldsymbol{\theta})$. This sequential approach can further increase simulation efficiency by orders of magnitude compared with the amortized counterpart (Papamakarios & Murray 2016), at the expense of forgoing observation-agnostic flexibility. Truncation schemes (Miller et al. 2020, 2021; Deistler, Goncalves & Macke 2022; Karchev, Trotta & Weniger 2022) also follow this sequential approach, by truncating the prior distribution at each sequential step in order to reduce the total number of simulations.

We briefly review the three main approaches to neural density estimation (Cranmer et al. 2020; see also Lueckmann et al. 2021, for a benchmark of the various algorithms). When referring to variants of these implementations we follow the nomenclature of Durkan, Murray & Papamakarios (2020).

3.1. Neural posterior estimation

Neural posterior estimation (NPE) was first introduced by Papamakarios & Murray (2016) and involves training a conditional density estimator to approximate the posterior density, such that $q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\boldsymbol{d}) \to p(\boldsymbol{\theta}|\boldsymbol{d})$, by minimizing the loss function

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{p(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\left[-\log q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\boldsymbol{d})\right]. \tag{9}$$

For sequential NPE, iteratively updating the proposal distribution between inference rounds results in the density estimator learning a proposal posterior density $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{d})$ that is related to the true posterior density by

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{d}) \propto \frac{\tilde{p}(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{d}). \tag{10}$$

Three variants of NPE have been introduced to recover the true posterior from the proposal posterior.

The original NPE method (NPE-A, Papamakarios & Murray 2016) trains a mixture density network to target the posterior distribution. A post-hoc analytical correction is then applied to the resulting proposal posterior to recover an approximation of the true posterior (cf. equation 10). NPE-A considers Gaussian or Gaussian mixture proposal distributions so that the correction factor can be computed analytically.

To circumvent the requirement for analytical computation, Lueckmann et al. (2017) propose a method (NPE-B) where the proposal correction is embedded as an importance weight in the loss function. Whilst more flexible than NPE-A, this method has been shown to suffer poor performance as the importance weights in the loss function are susceptible to high variance, resulting in early termination of training.

Finally, Greenberg, Nonnenmacher & Macke (2019) propose a NPE method (NPE-C) that reparametrizes the problem to recover a learned approximation $q_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\boldsymbol{d})$ of the true posterior from a density estimator $\tilde{q}_{\boldsymbol{\phi}}(\boldsymbol{\theta}|\boldsymbol{d})$ of the proposal posterior using a tractable sum of discrete atomic proposals over the support of the true posterior. This latter approach allows more flexibility in the choice of density estimator, including cutting-edge normalizing flow models (Papamakarios, Pavlakou & Murray 2017; Durkan et al. 2019). In our subsequent experiments we only consider this NPE method, which we simply refer to as NPE for the remainder of this article.

NPE learns the posterior density directly, typically for a probabilistic model from which samples can be drawn directly. For example, samples can be drawn directly from a Gaussian mixture density network or from a normalizing flow, where for the latter samples are first drawn from a simple base distribution such as a Gaussian and pushed forward through the flow to yield samples of the target distribution. Consequently, generating samples avoids the need for Markov chain Monte Carlo (MCMC) sampling and can be performed rapidly and in parallel, significantly reducing computation time for inference.

3.2. Neural likelihood estimation

Neural likelihood estimation (NLE) was first introduced by Papamakarios, Sterratt & Murray (2019) and involves training a conditional density estimator to approximate the likelihood function (considering it as a probability distribution over the data), such that $q_{\boldsymbol{\phi}}(\boldsymbol{d}|\boldsymbol{\theta}) \to p(\boldsymbol{d}|\boldsymbol{\theta})$, by minimizing the loss function

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{p(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\left[-\log q_{\boldsymbol{\phi}}(\boldsymbol{d}|\boldsymbol{\theta})\right]. \tag{11}$$

In contrast to NPE, sequential NLE can be implemented without a correction between $\tilde{q}_{\boldsymbol{\phi}}(\boldsymbol{d}|\boldsymbol{\theta})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{d}|\boldsymbol{\theta})$. In principle, simulations can be generated for any proposal distribution and, consequently, simulations from all sequential rounds, not just the latest, can be used when training (Papamakarios et al. 2019).

This ability to seamlessly optimize simulation efficiency, however, comes at the expense of requiring an external MCMC sampling stage to generate samples from the surrogate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ for inference, which increases inference time and computational cost significantly relative to NPE, where samples can be generated directly.

3.3. Neural ratio estimation

Neural ratio estimation (NRE) was first introduced by Hermans, Begy & Louppe (2020) and involves approximating the posterior density $p(\boldsymbol{\theta}|\boldsymbol{d})$ indirectly by learning a density ratio $r_{\boldsymbol{\phi}}(\boldsymbol{d}, \boldsymbol{\theta})$ that is proportional to the likelihood, where $\boldsymbol{\phi}$ denotes the model weights.

This is done by training a binary classifier to discriminate samples drawn from the joint and marginal distributions of training pairs. The classifier then learns the ratio

$$r_{\boldsymbol{\phi}}(\boldsymbol{d},\boldsymbol{\theta}) = \frac{p(\boldsymbol{d},\boldsymbol{\theta})}{p(\boldsymbol{d})p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{d}|\boldsymbol{\theta})}{p(\boldsymbol{d})} = \frac{p(\boldsymbol{\theta}|\boldsymbol{d})}{p(\boldsymbol{\theta})}. \tag{12}$$

A further NRE variant was devised by Durkan et al. (2020). This variant extends the binary classifier to a multiclass one, hence we adopt this variant for our numerical experiments and simply refer to it as NRE for the remainder of this paper.

Similarly to NLE, an additional MCMC sampling step can be used to generate samples from the surrogate posterior $r_{\boldsymbol{\phi}}(\boldsymbol{d},\boldsymbol{\theta})p(\boldsymbol{\theta})$. As with NLE, this increases inference time and computation cost significantly relative to NPE. Alternatively, one can sample from the prior when it is tractable and incoporate approximate importance sampling weights (given by the ratio itself).

4. BAYESIAN MODEL COMPARISON FOR SBI

We discussed the importance and challenge of computing the model evidence for Bayesian model selection in Section 2, which is a fundamental component of many scientific analyses. Separately, in Section 3, we discussed three recent neural density estimation techniques for parameter estimation in an SBI setting, which offer great promise for scientific analyses where the likelihood is often intractable or too costly to be evaluated. In this section, we unify these two topics by introducing a methodology to compute the Bayesian model evidence in all of the three neural density estimation approaches. The evidence computation technique corresponding to each neural density estimation approach is represented schematically in Fig. 1. Our approaches support density estimation training phases run in either an amortized or sequential setting.

4.1. Neural posterior estimation

Our approach to compute the evidence for NPE is shown schematically in the left-hand panel of Fig. 1. We use NPE to learn an approximation $q_{\boldsymbol{\psi}}^{\mathrm{NPE}}(\boldsymbol{\theta}|\boldsymbol{d})$ of the posterior, parametrized by network weights $\boldsymbol{\psi}$. This approach provides the ability to rapidly generate samples directly from the surrogate posterior, i.e. $\boldsymbol{\theta}_i \overset{\text{direct}}{\sim} q_{\boldsymbol{\psi}}^{\mathrm{NPE}}(\boldsymbol{\theta}|\boldsymbol{d})$. While NPE also provides the ability to evaluate the surrogate normalized posterior, the normalization constant itself, i.e. the model evidence, is not accessible. To compute the model evidence we therefore adopt the learned harmonic mean estimator, using the samples drawn directly from the surrogate posterior. For the learned harmonic mean estimator it is also necessary to evaluate the likelihood at sample positions (see equation 7), hence we adopt NLE to provide a surrogate likelihood. Using NLE we learn an approximation $q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta})$ of the likelihood, parametrized by a separate set of network weights $\boldsymbol{\phi}$. With a set of posterior samples and the surrogate likelihood learned by NLE to hand, we use the learned harmonic mean estimator to obtain an estimate of the reciprocal of the model evidence by

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\theta}_i)}{q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\text{direct}}{\sim} q_{\boldsymbol{\psi}}^{\mathrm{NPE}}(\boldsymbol{\theta}|\boldsymbol{d}). \tag{13}$$

The proposed approach to compute the model evidence in the NPE setting does involve training two neural density estimators, both NPE and NLE. However, it does not require any MCMC sampling. Samples can be generated directly from the surrogate posterior learned by NPE (e.g. by pushing samples from a simple base distribution such as a Gaussian through a normalizing flow), which is highly efficient and can also be computed in parallel.

Given trained NPE and NLE surrogate densities, an alternative naïve technique can also be considered to estimate the model evidence. For any model parameter $\boldsymbol{\theta}$ the ratio of the unnormalized surrogate posterior, formed from the surrogate likelihood and prior, to the normalized surrogate posterior, i.e.

$$\frac{q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q_{\boldsymbol{\psi}}^{\mathrm{NPE}}(\boldsymbol{\theta}|\boldsymbol{d})}, \tag{14}$$

provides an estimate of the evidence. An estimate of the evidence is thus recovered for a single parameter $\boldsymbol{\theta}$, which need not be drawn from any particular distribution. However, such an estimate of the evidence will be incredibly noisy, i.e. will have an extremely large variance. Many parameters can be used to generate many estimates of the evidence that can be averaged. Nevertheless, the resulting estimate of the evidence remains highly noisy with a very large variance. This naïve estimator relies on the ratio of two approximate quantities, hence approximation errors compound. Contrast this with the learned harmonic mean estimator. While our learned harmonic mean approach does use NPE to learn a surrogate posterior $q_{\boldsymbol{\psi}}^{\mathrm{NPE}}(\boldsymbol{\theta}|\boldsymbol{d})$, the density is never evaluated. We only require samples from the corresponding distribution. The learned harmonic mean does require learning the importance target $\varphi(\boldsymbol{\theta})$, and this is indeed learned to approximate the posterior, but the target need only be normalized and have tighter tails than the true posterior $p(\boldsymbol{\theta}|\boldsymbol{d})$ – it does not need to be an accurate approximation of the posterior. Consequently, our proposed approach to compute the evidence in the NPE setting, based on the learned harmonic mean estimator, does not suffer compounding sources of error and thus provides increased stability over the naïve approach.

While the focus of the current article is SBI, we also comment that the ideas presented here can also be applied to accelerate evidence computation for likelihood-based inference. Crucially, throughout our approach to compute the evidence in the NPE setting, MCMC sampling is not required. Posterior samples can be generated directly, rapidly and in parallel. If a likelihood is available this can simply be substituted for the surrogate likelihood learned by NLE. Therefore in the likelihood-based setting the approach can be altered to leverage the speed of posterior sample generation of NPE, while adopting the analytical likelihood function, to obtain a rapid estimate of the reciprocal evidence without any further computation, as described by

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\theta}_i)}{p(\boldsymbol{d}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\text{direct}}{\sim} q_{\boldsymbol{\psi}}^{\mathrm{NPE}}(\boldsymbol{\theta}|\boldsymbol{d}). \tag{15}$$

Clearly in this setting NLE need not be performed.

4.2. Neural likelihood estimation

Our approach to compute the evidence for NLE is shown schematically in the central panel of Fig. 1. We use NLE to learn an approximation of the likelihood function $q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta})$, parametrized by network weights $\boldsymbol{\phi}$. As is typical for NLE, this approach requires MCMC sampling to generate samples from the unnormalized surrogate posterior, i.e. $\boldsymbol{\theta}_i \overset{\text{MCMC}}{\sim} q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. NLE also provides the ability to evaluate the surrogate likelihood. With a set of posterior samples and the surrogate likelihood learned by NLE to hand, we use the learned harmonic mean estimator to compute an estimate of the reciprocal of the model evidence by

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\theta}_i)}{q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\text{MCMC}}{\sim} q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{16}$$

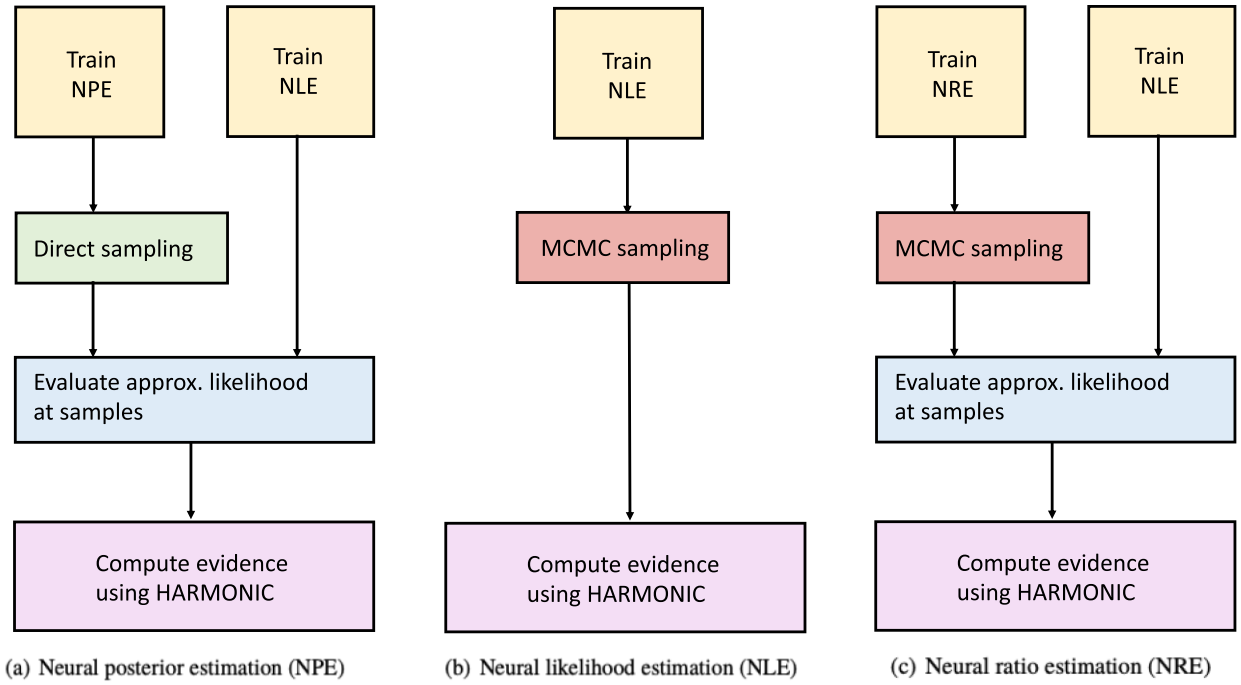

Figure 1. Schematic overview of three novel techniques that we introduce to compute the evidence in SBI settings for neural posterior, likelihood, and ratio estimation methods (NPE, NLE, NRE, respectively). The *top* training blocks represent all phases of neural density estimation, where each block can be run in an amortized or sequential setting.

This proposed approach to compute the model evidence in the NLE setting involves training only one neural density estimator, which is decidedly more efficient than training two such estimators as required in the NPE and NRE settings (cf. Sections 4.1 and 4.3). However, it does require MCMC sampling to generate samples from the unnormalized surrogate posterior which is required to compute the evidence using the learned harmonic mean estimator.

With a trained NLE surrogate density $q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta}_i)$ and the prior $p(\boldsymbol{\theta})$ to hand, alternative techniques that only require the likelihood function and prior could also be considered to compute an estimate of the evidence.

4.3. Neural ratio estimation

Our approach to compute the evidence for NRE is shown schematically in the right-hand panel of Fig. 1. We use NRE to indirectly learn an approximation $r_{\boldsymbol{\psi}}^{\mathrm{NRE}}(\boldsymbol{d},\boldsymbol{\theta})p(\boldsymbol{\theta})$ of the posterior, parametrized by network weights $\boldsymbol{\psi}$. Similarly to our NPE approach, the normalization constant of the surrogate posterior learned by NRE, i.e. the model evidence, is not accessible. We therefore adopt the learned harmonic mean estimator to compute the model evidence, for which it is also necessary to evaluate the likelihood at sample positions (see equation 7), hence we adopt NLE to provide a surrogate likelihood. Using NLE we learn an approximation $q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta})$ of the likelihood, parametrized by network weights $\boldsymbol{\phi}$. With a set of posterior samples and the surrogate likelihood learned by NLE to hand, we use the learned harmonic mean estimator to obtain an estimate of the reciprocal of the model evidence by

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N}\frac{\varphi(\boldsymbol{\theta}_i)}{q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\mathrm{MCMC}}{\sim} r_{\boldsymbol{\psi}}^{\mathrm{NRE}}(\boldsymbol{d},\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{17}$$

The proposed approach to compute the model evidence in the NRE setting does involve training two neural density estimators, both NRE and NLE. Furthermore, external MCMC sampling is required above to generate samples from the trained NRE surrogate posterior.

Alternatively, since NRE provides access to an approximation of the normalized posterior by $r_{\boldsymbol{\psi}}^{\mathrm{NRE}}(\boldsymbol{d},\boldsymbol{\theta})p(\boldsymbol{\theta})$, via importance sampling one could instead sample from the prior to avoid the need for MCMC sampling:

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N}\frac{r_{\boldsymbol{\psi}}^{\mathrm{NRE}}(\boldsymbol{d},\boldsymbol{\theta}_i)\varphi(\boldsymbol{\theta}_i)}{q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\mathrm{direct}}{\sim} p(\boldsymbol{\theta}). \tag{18}$$

However, the above estimator involves the ratio of two approximate quantities and so approximation errors compound. Furthermore, one could consider using the NRE approximation of the normalized posterior for the learned harmonic mean target distribution $\varphi(\boldsymbol{\theta})$:

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left[r_{\boldsymbol{\psi}}^{\mathrm{NRE}}(\boldsymbol{d},\boldsymbol{\theta}_i)\right]^2}{q_{\boldsymbol{\phi}}^{\mathrm{NLE}}(\boldsymbol{d}|\boldsymbol{\theta}_i)}, \quad \boldsymbol{\theta}_i \overset{\mathrm{direct}}{\sim} p(\boldsymbol{\theta}). \tag{19}$$

However, such an estimator is also unlikely to be well-behaved since we have not explicitly ensured the tails of $\varphi(\boldsymbol{\theta})$ are narrower than the posterior and it is a ratio of two approximate quantities, one of which is squared, and so approximation errors compound. We therefore do not consider these variants of the estimator further.

An alternative way to compute the Bayes factor B_{12} between two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, i.e. the ratio of model evidences (equation 4), in the NRE setting is to train an additional NRE model as a binary classifier to discriminate samples from the joint and marginal distribution of the two models, respectively. The classifier then learns the ratio

$$r_{\boldsymbol{\psi}_{12}}(\boldsymbol{d},\boldsymbol{\theta}) = \frac{p(\boldsymbol{d},\boldsymbol{\theta}|\mathcal{M}_1)}{p(\boldsymbol{d}|\mathcal{M}_2)p(\boldsymbol{\theta}|\mathcal{M}_2)}, \tag{20}$$

where $\boldsymbol{\psi}_{12}$ denotes the network weights for a model trained in such a manner. Following similar notation, the standard neural ratio for a

single model, say model $\mathcal{M}_1$, can be denoted $r_{\psi_{11}}$. While it is not possible to estimate the evidence of a single model directly, the Bayes factor comparing the two models, which is the critical quantity for model comparison, can then be recovered by

$$B_{12} = \frac{r_{\psi_{12}}(\boldsymbol{d}, \boldsymbol{\theta})}{r_{\psi_{11}}(\boldsymbol{d}, \boldsymbol{\theta})} \frac{p(\boldsymbol{\theta}|\mathcal{M}_2)}{p(\boldsymbol{\theta}|\mathcal{M}_1)}. \tag{21}$$

We understand this method is already known to the SBI community but were not able to locate any references discussing or applying it. Since this approach does not lie within the family of methodologies introduced in the current article, which leverage the learned harmonic mean estimate to compute the model evidence from samples of the surrogate posterior distribution, we leave the analysis of this approach to further work.

5. NUMERICAL EXPERIMENTS

Here, we present the results from our numerical experiments that demonstrate and validate our SBI evidence calculation techniques. For validation purposes, for each problem, we compare the value of the evidence computed by our proposed approach (which we stress does *not* include any knowledge of the likelihood) to values computed by likelihood-based approaches, either derived analytically (when possible) and/or computed numerically by likelihood-based algorithms (e.g. by HARMONIC, MULTINEST, and/or POLYCHORD). Of course in practical SBI settings, likelihood-based alternatives will not typically be available. Nevertheless, it is useful to consider problems here where the likelihood is available so that we can validate our SBI evidence computation techniques against likelihood-based alternatives. All of the SBI examples were implemented using the SBI[8] software (Tejero-Cantero et al. 2020).

5.1. Linear Gaussian

The first problem we consider is that of a simple simulator which linearly adds Gaussian noise ϵ_i to the value of the parameters θ_i, for an arbitrary number of parameters $i = 1, \ldots N$:

$$\mathrm{d}_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1). \tag{22}$$

This is a standard test problem in the SBI software, which we trivially generalize to arbitrary dimension N. The Gaussian noise has zero mean and unit variance, and for the model parameters we assume a uniform prior $\theta_i \sim \mathcal{U}[-2, 2]$ for each component i. The likelihood for this model is Gaussian in the parameters $\boldsymbol{\theta} = \{\theta_1, \ldots \theta_N\}$:

$$p(\boldsymbol{d}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{3/2}} \exp\left(-\frac{(\boldsymbol{d} - \boldsymbol{\theta})^2}{2}\right). \tag{23}$$

We assume an observation $\boldsymbol{d}_0 = (0, 0, 0)$. For this model the Bayesian evidence can be computed analytically:

$$\begin{aligned} z &= \frac{1}{4^N\ (2\pi)^{N/2}} \int_{-2}^{2} \mathrm{d}\theta_1 \ldots \int_{-2}^{2} \mathrm{d}\theta_N\ \exp\left(-\frac{\theta_1^2 + \cdots + \theta_N^2}{2}\right) \\ &= \frac{\left[\mathrm{erf}\left(\sqrt{2}\right)\right]^N}{4^N}. \end{aligned} \tag{24}$$

Fig. 2 summarizes the results obtained from our model evidence estimates for the linear Gaussian problem. We consider the three cases $N = \{3, 10, 20\}$, shown in the upper, middle, and bottom panels, respectively, to investigate any dependence of our evidence estimates on the number of parameters considered. In all panels, the *pink* background section shows results for likelihood-based methods for validation purposes, while results for SBI (likelihood-free) methods are shown in the *light brown* region. For all methods we repeat the evidence estimation exercise 25 times to empirically describe the statistical distributions of the model evidence estimates, shown by the blue areas in each 'violin' of Fig. 2. All of the evidence values reported in Fig. 2 can be compared with the analytical value of equation (24), overplotted by the *red* dashed line.

Likelihood-based approaches include: (1) MULTINEST, which produces samples and evidence estimates, run with importance sampling (Feroz et al. 2019), 1000 live points, efficiency sampling of 0.3 and evidence tolerance of 0.01, resulting in $\sim 10^4$ samples; (2) POLYCHORD, which also produces samples as well as evidence estimates, run using 1000 live points, resulting in $\sim 10^4$ samples; (3) HARMONIC, which produces evidence estimates from posterior samples, thus we adopt the affine sampler of Goodman & Weare (2010), implemented in the EMCEE[9] software (Foreman-Mackey et al. 2013), to generate 10^5 post burn-in posterior samples from 100 random walkers and adopt a hypersphere model for the learned importance target $\varphi(\boldsymbol{\theta})$, with radius equal to the square root of the number of parameters. The number of samples generated by MULTINEST and POLYCHORD is dynamic, depending on, e.g. the tolerance parameter, whereas for EMCEE sampling for HARMONIC we adopt a conservative, fixed number of samples, since this is the configuration we will also use when considering the SBI scenarios. Consequently, the variances of the nested sampling approaches should not be compared directly to those of the learned harmonic mean due to the differing number of samples.

We analyse the performance of the estimators in terms of their bias and variance (also adopting the terminology *accuracy* and *precision*, respectively, as common in the astrophysical literature). All of the three likelihood-based methods provide unbiased average estimates of the evidence. As mentioned, the variance of these estimators should not be directly compared due to the differences in the number of samples (an analysis of the performance of nested sampling and learned harmonic mean approaches would be interesting but here we are focused on validating our proposed techniques to compute the evidence for SBI scenarios).

The results for the SBI methods are shown in the *light brown* section of the plot. We report results for NPE, NLE, and NRE, and for each of them we provide an estimate using the amortized as well as the sequential approach. For all methods we use 10^5 simulations in the amortized approach, while in the sequential one we use 10 rounds with 10^4 simulations each, thus totalling the same number of simulations for the two approaches. For each SBI method after training a density estimator (in an amortized or sequential fashion) we collect a total of 10^5 posterior samples, either directly for NPE or by MCMC sampling using 100 EMCEE random walkers for NLE and NRE. We train the HARMONIC importance target model using 20 per cent of the samples and use the remaining 80 per cent to compute the evidence. As explained in Sections 4.1 and 4.3, NPE and NRE require an additional training of an NLE estimator to provide a surrogate likelihood.

All of the SBI evidence computation techniques provide estimates of the evidence whose distribution captures the true analytic evidence in the $N = 3$ and $N = 20$ cases, although there is a residual bias in many cases. Moreover, for $N = 10$, the NPE estimator distribution does not always capture the true evidence. The fact that the estimates

[8] https://github.com/mackelab/sbi

[9] https://github.com/dfm/emcee

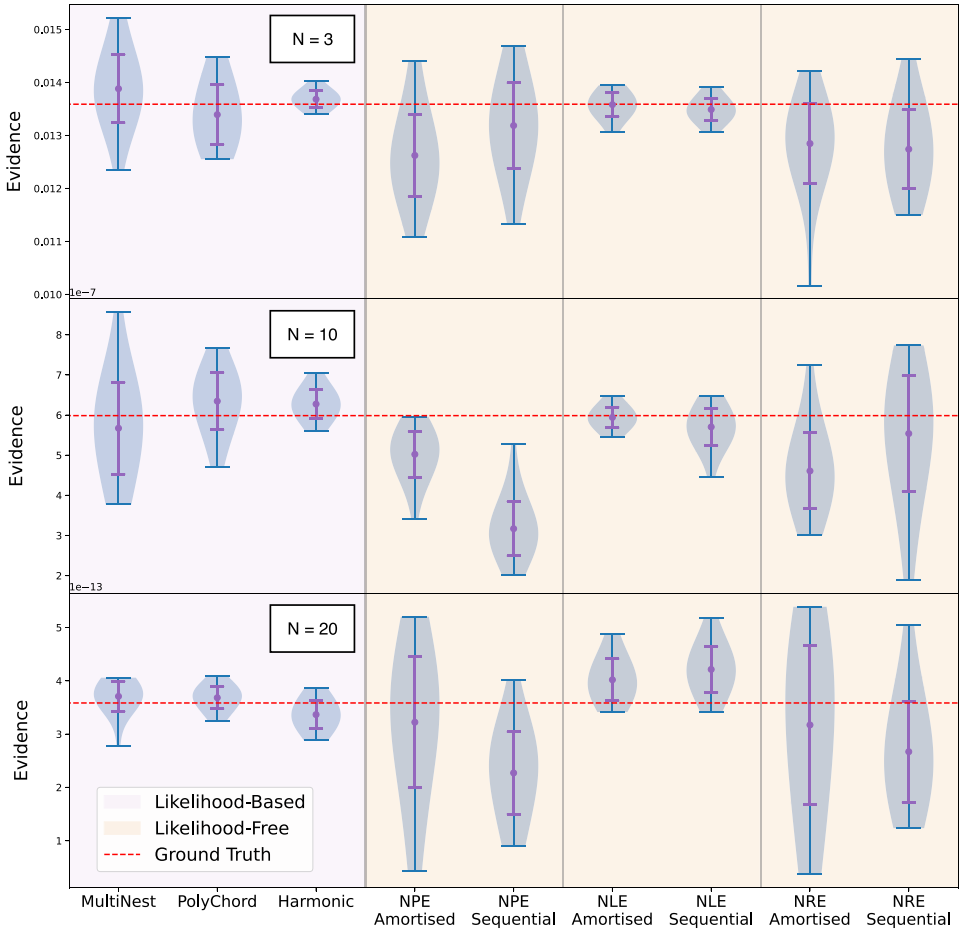

Figure 2. Model evidence values estimated with different likelihood-based and simulation-based (likelihood-free) methods for the linear Gaussian example described in Section 5.1, whose analytical truth value is shown by the *red* dashed line. We consider three different dimensions, namely $N = \{3, 10, 20\}$, shown in the upper, middle, and bottom panel, respectively. For all methods we repeat the evidence estimation exercise 25 times to empirically describe the statistical distributions of the model evidence estimates, shown by the blue areas in each 'violin'. The mean and one standard deviation error bars are illustrated in purple. The *pink* section of this plot shows likelihood-based results obtained with MULTINEST, POLYCHORD, and HARMONIC (the latter using samples from EMCEE). Note that the variance of the evidence estimates obtained by these three methods in the likelihood-based setting should be compared directly due to differing numbers of samples (see Section 5.1 for a detailed discussion). The *light brown* section of the plot shows results for the simulation-based evidence pipelines summarized in Fig. 1. These are all based on the use of HARMONIC to derive evidence estimates from posterior samples obtained with NPE, NLE, and NRE, in their amortized and sequential variants.

for dimension $N = 20$ – the more challenging setting – do encompass the true evidence suggests that these biases may be due to insufficient training of the underlying SBI models and/or their difficulty to scale to higher dimensions. The variances of the SBI estimates are generally larger than the likelihood-based approaches, which is to be expected since in contrast to the likelihood-based setting we do not include any knowledge of the likelihood. We also note that each evidence estimate with an SBI method does include some training noise due to the fact that we repeat the training at every iteration. The NLE approaches exhibit in general less bias than the NPE and NRE estimators. This may be due to the fact that the NLE approach requires only a single neural density estimator (whereas the NPE and NRE approaches require two), resulting in fewer sources of approximation error.

In Fig. 3, we investigate the dependence of the evidence estimate on the total number of simulations used in training the neural density models. We consider the case $N = 3$ and we report the results for the amortized and sequential approach to NPE, NLE, and NRE, varying the number of simulations from 10^4, to 10^5, to 10^6. We observe that in all cases that as the number of simulations increases the evidence estimates are less biased and the variances are reduced. This initial analysis suggests that evidence values can be computed accurately in SBI settings, although care should be taken to ensure a sufficient number of simulations are used. A more extensive analysis of the accuracy and precision of SBI approaches for evidence calculation would be welcome, along the lines of the extensive and informative study performed by Hermans et al. (2022). However, such an analysis requires substantial computational resources and is beyond the scope of this article, where we introduce these new methodologies.

5.2. Radiata pine

The second problem we considered is one of the classic benchmark examples used to evaluate techniques to estimate the model evidence (Friel & Wyse 2012; Enderlein 1961). We refer to McEwen et al. (2021), who demonstrated the effectiveness of the learned harmonic mean estimator for this example, for a more detailed presentation of

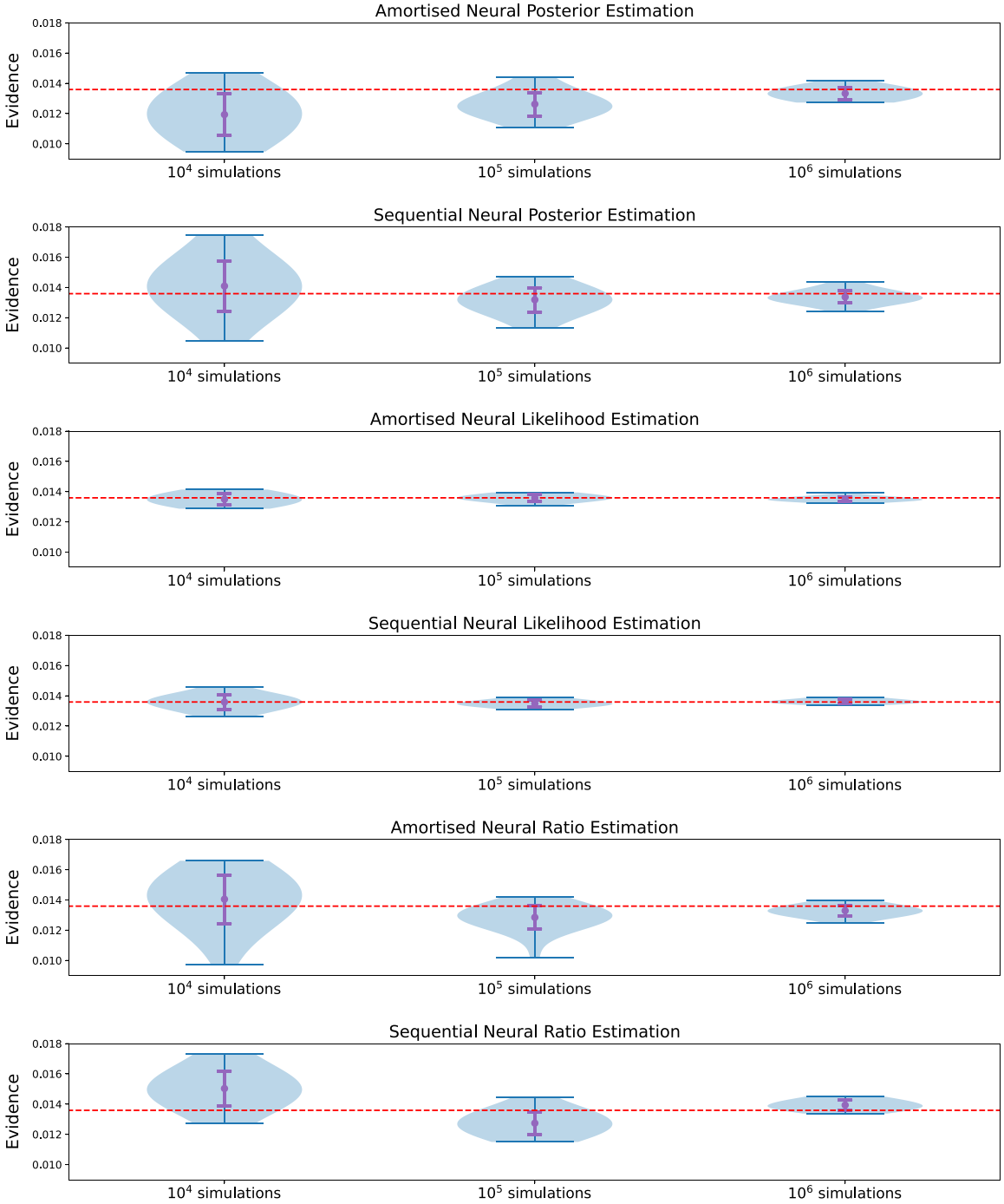

Figure 3. Dependence of the accuracy of the evidence estimates on the number of simulations used for training with simulation-based methods for the three-dimensional linear Gaussian example, whose analytical truth value is shown by the *red* dashed line. The mean and one standard deviation error bars are illustrated in purple. We report results for the amortized and sequential versions of NPE, NLE, and NRE. The evidence pipelines are summarized in Fig. 1 and are all based on the use of HARMONIC to derive evidence estimates. In all cases as the number of simulations increases the evidence estimates are less biased and the variances are reduced. These results suggest that evidence values can be computed accurately in SBI settings, although care should be taken to ensure a sufficient number of simulations are used.

the problem: here we simply report the main points relevant to our evidence estimation task.

Our data set is comprised of measurements of Radiata pine trees of the maximum compression strength parallel to the grain y_i, for $i = 1 \ldots 42$. The original scientific problem can be stated in terms of two models: in Model 1 the density x_i is assumed as a predictor for y_i, while in Model 2 the predictor is assumed to be the resin-adjusted density z_i. Both predictors are modelled with a Gaussian linear regression model, for which the value of the evidence can be derived analytically for each model. For brevity, we report results only for Model 1; calculations for Model 2 are identical (we did also experiment with this second model, finding excellent agreement with the analytical estimates of the evidence). In Model 1, denoting with $\bar{x} = \frac{1}{n}\sum_{i=1}^{42} x_i$ the average density across the trees specimens, the

maximum compression strength y_i is given by

$$y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \tau^{-1}). \tag{25}$$

The model parameters are $\{\alpha, \beta, \tau\}$, whose prior distributions are

$$\alpha \sim \mathcal{N}(\mu_\alpha, (r_0\tau)^{-1}), \ \beta \sim \mathcal{N}(\mu_\beta, (s_0\tau)^{-1}), \ \tau \sim \mathrm{Ga}(a_0, b_0), \tag{26}$$

with $(\mu_\alpha, \mu_\beta, r_0, s_0, a_0, b_0) = (3000, 185, 0.06, 6, 3, 2 \times 300^2)$. The evidence for this model can be computed analytically (cf. McEwen et al. 2021, equation 104); the numerical value of its logarithm is $\log z = -310.12829$.

Fig. 4 summarizes our findings for the Radiata pine example; the colour codes are the same as in Fig. 2. We repeat the same experiments as in the linear Gaussian example, except this time for simplicity we do not attempt to calculate the evidence with MULTINEST or POLYCHORD (as this would require some effort to adapt the prior function for the Radiata pine model to be compatible with the uniform distribution on the unit cube required by these nested samplers). Therefore, for the likelihood-based case, we report only numerical results obtained by applying HARMONIC to EMCEE samples, using a kernel density estimate for the learned harmonic mean estimator importance target, with radius 0.02 of the target distribution. As in the linear Gaussian example, we use 20 per cent of 10^5 EMCEE posterior samples from 100 random walkers to train HARMONIC, and compute an estimate of the evidence with the remaining 80 per cent. As we can see in the *pink* section of the plot, this provides unbiased and tight estimates of the evidence.

In the *light brown* background section of Fig. 4 we can compare results for SBI methods. The number of simulations we use to train the density estimators in the various methods is the same as in the baseline case for the linear Gaussian example, as is the number of posterior samples used to train HARMONIC and derive evidence estimates. All of the SBI evidence pipelines provide reasonably accurate estimates of the evidence, with distribution ranges capturing the true analytic evidence. The NPE and NRE approaches again exhibit some bias, which is nevertheless within the spread of the distribution of values. The NLE estimates again show good agreement with the reference values.

5.3. Gravitational waves

The final problem we consider is a simulated measurement of a gravitational wave (GW) signal from a single interferometer. We consider a merger between two black holes of mass $M_1 = M_2 = 20\ \mathrm{M}_\odot$, following a similar numerical set-up to the one considered by Jeffrey & Wandelt (2020). We compute the noiseless time series of the strain signal using the PYCBC[10] software (Biwer et al. 2019). We only consider the '+' polarization of the detector strain $h_{+,\times}$. The duration of the signal is ∼0.12 s, sampled at steps of duration ∼488 μs each. We rescale the signal by a multiplicative factor 10^{21} in order to work with values $\mathcal{O}(1)$. For each point in parameter space, our simulated observable is a noisy GW form obtained by adding Gaussian noise with zero mean and standard deviation $\sigma = 0.3$ to the noiseless template from PYCBC.

For this inference problem we vary the two black hole masses M_1 and M_2, each one over a uniform prior $\mathcal{U}[10, 30]\,\mathrm{M}_\odot$. We do not consider geometrical properties of the black holes such as spin or inclination angle, similarly to Jeffrey & Wandelt (2020). As noted by Hermans et al. (2022), who studied a similar GW SBI problem, such a simulated experimental set-up requires significant computational demand. Hence, for this example, we run only 10 repetitions of each inference method to empirically describe the statistical distribution of evidence estimates. These estimates mimic a realistic scenario within GW data analysis pipelines comparing model evidences for two different numerical approximant models assumed in the generation of the noiseless template waveforms. The first of these two waveform models corresponds to the actual one used to generate the simulated observation, a reduced-order effective-one-body model (SEOBNR, Taracchini et al. 2014), which we refer as the *source* model. The second model we consider is an inspiral-merger-ringdown phenomenological model (Hannam et al. 2014), which we refer to as the *alternative* model. For this configuration, we expect to find the Bayes factors comparing models to favour the source SEOBNR model. We verify this numerically, taking advantage of this problem's deliberately low-dimensional parameter space, which allows us to compute an estimate of the evidence for each model using direct numerical integration. We find the logarithm of the Bayes factor computed by direct numerical integration to be ∼3.25, favouring the source model as anticipated.

We run the inference pipeline in multiple likelihood-based contexts, namely (1) obtaining samples and evidence estimates with POLYCHORD, using the same configuration for this method as the one described in Section 5.1 and (2) sampling the parameter space with EMCEE, collecting the same number of posterior samples as in both previous numerical examples, and using HARMONIC to obtain an estimate of the evidence, with a kernel density estimate for the importance target distribution of radius 0.002 and 0.02 for the source and alternative models, respectively. Numerical results from these likelihood-based methods are reported in the *pink* background section of Fig. 5, where we show violin plots for log-Bayes factors computed with POLYCHORD and HARMONIC. The distribution of log-Bayes factors always favours the true underlying source model and, as in previous examples, we find strong agreement between each likelihood-based approach.

Numerical results from the SBI evidence estimates are shown in the *light brown* background section of Fig. 5. The evidence pipelines we consider are the same as described in Section 4 and for each pipeline, the number of simulations used to train the density estimators are the same as those used for the baseline linear Gaussian and Radiata pine examples in Sections 5.1 and 5.2, for both amortized and sequential approaches. We also use the same number of posterior samples to train HARMONIC and derive estimates of the evidence. We notice that all SBI methods produce log-Bayes factor estimates in general agreement with the likelihood-based ones, albeit with a comparatively larger variance than presented in Figs 2 and 4, due to compounding errors from the multiple evidence estimates required to obtain Bayes factors (cf. equation 4). NLE produces more unbiased estimates of the log-Bayes factors compared with NPE and NRE – a similar trend to that observed in the linear Gaussian and Radiata pine examples (cf. Figs 2 and 4).

6. CONCLUSIONS

In this article, we propose a novel methodology to compute the model evidence for modern neural density estimation approaches to SBI using the learned harmonic mean estimator. Our approach leverages the property of the learned harmonic mean estimator that it is decoupled from the sampling strategy and only requires samples of the posterior. This allows us to develop SBI model comparison techniques for all three main neural density estimation approaches: NPE, NLE, and NRE.

[10] https://github.com/gwastro/pycbc

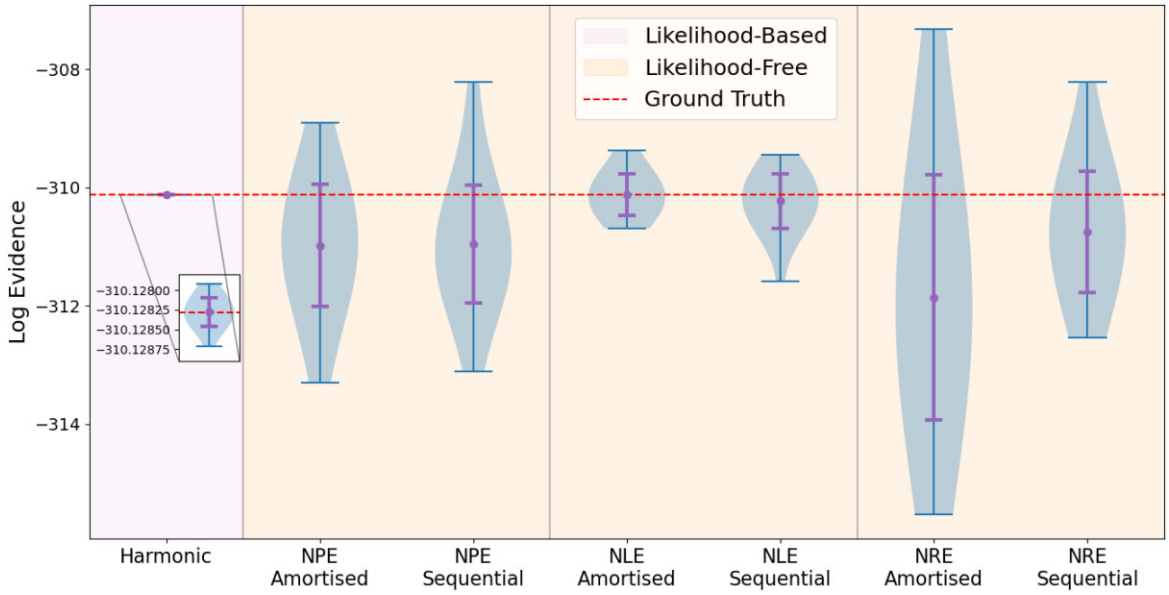

Figure 4. Model evidence values estimated with different likelihood-based and simulation-based (likelihood-free) methods for the Radiata pine example described in Section 5.2, whose analytical truth value is shown in *red*. Colour codes and labels are consistent with Fig. 2.

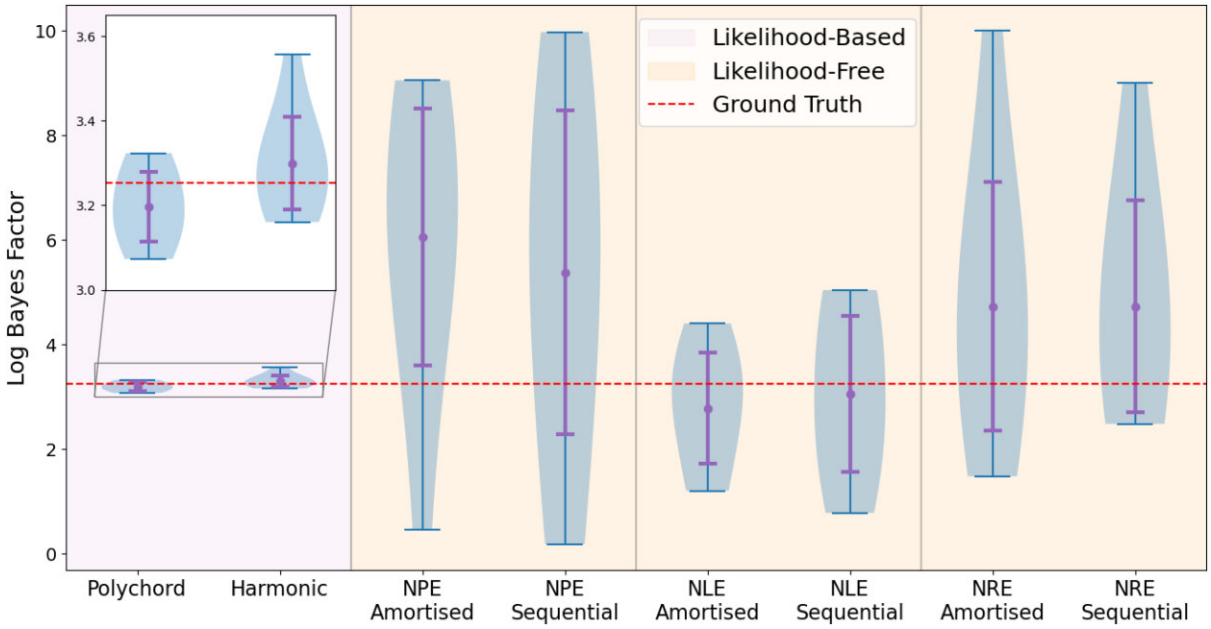

Figure 5. Logarithm of the Bayes factors between source and alternative waveform models estimated with different likelihood-based and simulation-based (likelihood-free) methods for the GW example described in Section 5.3. Shown in *red*, an estimate of the true value of the logarithm of the Bayes factor was obtained using numerical integration. Colour codes and labels are consistent with Figs 2 and 4.

We demonstrate and validate our SBI evidence calculation techniques on a range of inference problems, using the learned harmonic mean estimator as implemented in the HARMONIC software. We validate all SBI evidence estimator approaches, computed using HARMONIC, against those computed by likelihood-based alternatives. We find that HARMONIC produces values of the evidence that are in excellent agreement with those computed by the likelihood-based nested sampling algorithms MULTINEST and POLYCHORD. Our results suggest that the learned harmonic mean estimator can be reliably used as an alternative to nested sampling for evidence estimation.

We also compare the different SBI evidence computation approaches that we propose and find that the NLE evidence estimation approach provides more accurate evidence estimates compared with the NPE and NRE approaches. This result is particularly encouraging for applications to cosmological scenarios, where it is very common to perform SBI inference using the PYDELFI software, which indeed implements (sequential) NLE to sample the posterior distribution.

Overall, our methodology and proof-of-concept analysis highlight the potential of the learned harmonic mean estimator as an additional tool for Bayesian model selection in SBI settings. Future research will focus on extending the applicability of HARMONIC to larger data

and parameter spaces, as well as on estimating the variance of the estimators directly, folding in both sampling variance and neural density approximate error. Application to more realistic examples and problems will be insightful, as well as a thorough analysis of the accuracy and precision of the proposed approach; we advocate a future study similar to Hermans et al. (2022) but focused on model selection. Our hope is that this article provides a first step towards the computation of the Bayesian model evidence in SBI scenarios, in order to facilitate principled and robust Bayesian model selection.

ACKNOWLEDGEMENTS

We thank T. Kitching and M. Zaldarriaga for useful discussions. ASM is supported by the MSSL STFC Consolidated Grant ST/W001136/1 and the Leverhulme Trust. MMD is supported by the Science and Technology Facilities Council (STFC) Centre for Doctoral Training (CDT) in Data Intensive Science (DIS) at UCL. MAP is supported by EPSRC grant EP/W007673/1. This work used computing facilities provided by the UCL Cosmoparticle Initiative and also facilities funded by the Research Capital Investment Fund (RCIF) provided by UKRI, and partially funded by the UCL Cosmoparticle Initiative and the UCL CDT in DIS.

DATA AVAILABILITY

HARMONIC is freely available at https://github.com/astro-informatics/harmonic.

REFERENCES

Alsing J., Charnock T., Feeney S., Wandelt B., 2019, MNRAS, 488, 4440
Ashton G. et al., 2022, Nature Rev. Methods Primers, 2, 39
Beaumont M. A., 2019, Annu. Rev. Stat. Appl., 6, 379
Beltrán M., García-Bellido J., Lesgourgues J., Liddle A. R., Slosar A., 2005, Phys. Rev. D, 71, 063532
Biwer C. M., Capano C. D., De S., Cabero M., Brown D. A., Nitz A. H., Raymond V., 2019, PASP, 131, 024503
Brewer B. J., Pártay L. B., Csányi G., 2011, Stat. Comput., 21, 649
Bridges M., Lasenby A., Hobson M., 2006, MNRAS, 369, 1123
Buchner J., 2023, Stat. Surv., 17, 169
Buchner J. et al., 2014, A&A, 564, A125
Cai X., McEwen J. D., Pereyra M., 2022, Stat. Comput., 32, 87
Chib S., 1995, J. Am. Stat. Assoc., 90, 1313
Cole A., Miller B. K., Witte S. J., Cai M. X., Grootes M. W., Nattino F., Weniger C., 2022, J. Cosmol. Astropart. Phys., 2022, 004
Cranmer K., Brehmer J., Louppe G., 2020, Proc. Natl. Acad. Sci., 117, 30055
Deistler M., Goncalves P. J., Macke J. H., 2022, in Koyejo S., Mohamed S., Agarwal A., Belgrave D., Cho K., Oh A., eds, Advances in Neural Information Processing Systems, Vol. 35. Curran Associates, Inc., NY, USA, p. 23135, available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/9278abf072b58caf21d48dd670b4c721-Paper-Conference.pdf
Durkan C., Bekasov A., Murray I., Papamakarios G., 2019, Proc. 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., p. 12
Durkan C., Murray I., Papamakarios G., 2020, Proc. 37th International Conference on Machine Learning. JMLR.org, p. 11
Enderlein G., 1961, Biom. Z., 3, 145
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2019, Open J. Astrophys., 2, 10
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Friel N., Wyse J., 2012, Stat. Neerl., 66, 288
Gelfand A. E., Dey D. K., 1994, J. R. Stat. Soc. B, 56, 501
Goodman J., Weare J., 2010, Commun. Appl. Math.Comput. Sci., 5, 65
Greenberg D., Nonnenmacher M., Macke J., 2019, in Chaudhuri K., Salakhutdinov R., eds, Proc. Mach. Learn. Res. Vol. 97, Proc. 36th International Conference on Machine Learning. PMLR, p. 2404. Available at: https://proceedings.mlr.press/v97/greenberg19a.html
Gregory P., 2005, Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica® Support. Cambridge Univ. Press, Cambridge
Handley W. J., Hobson M. P., Lasenby A. N., 2015a, MNRAS, 450, L61
Handley W. J., Hobson M. P., Lasenby A. N., 2015b, MNRAS, 453, 4385
Hannam M., Schmidt P., Bohé A., Haegel L., Husa S., Ohme F., Pratten G., Pürrer M., 2014, Phys. Rev. Lett., 113, 151101
Heavens A., Fantaye Y., Mootoovaloo A., Eggers H., Hosenie Z., Kroon S., Sellentin E., 2017, preprint (arXiv:1704.03472)
Hermans J., Begy V., Louppe G., 2020, in Daumé H.,, III, Singh A., eds, Proc. Mach. Learn. Res. Vol. 119, Proc. 37th International Conference on Machine Learning. PMLR, p. 4239 . Available at: https://proceedings.mlr.press/v119/hermans20a.html
Hermans J., Delaunoy A., Rozet F., Wehenkel A., Begy V., Louppe G., 2022, Trans. Mach. Learn. Res.
Jeffrey N., Wandelt B. D., 2020, preprint (arXiv:2011.05991)
Karchev K., Trotta R., Weniger C., 2022, MNRAS, 520, 1056
Lemos P. et al., 2023, MNRAS, 521, 1184
Llorente F., Martino L., Delgado D., Ló pez-Santiago J., 2023, SIAM Rev., 65, 3
Lueckmann J.-M., Gonçalves P. J., Bassetto G., Öcal K., Nonnenmacher M., Macke J. H., 2017, Proc. 31st International Conference on Neural Information Processing Systems. NIPS'17. Curran Associates, Inc., NY, USA, p. 1289
Lueckmann J.-M., Boelts J., Greenberg D., Goncalves P., Macke J., 2021, in Banerjee A., Fukumizu K., eds, Proc. Mach. Learn. Res. Vol. 130, Proc. 24th International Conference on Artificial Intelligence and Statistics. PMLR, p. 343. Available at: https://proceedings.mlr.press/v130/lueckmann21a.html
McEwen J. D., Wallis C. G. R., Price M. A., Docherty M. M., 2021, preprint (arXiv:2111.12720)
Miller B. K., Cole A., Louppe G., Weniger C., 2020, preprint (arXiv:2011.13951)
Miller B. K., Cole A., Forré P., Louppe G., Weniger C., 2021, in Ranzato M., Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W., eds, Advances in Neural Information Processing Systems, Vol. 34. Curran Associates, Inc., NY, USA, p. 129. Available at: https://proceedings.neurips.cc/paper_files/paper/2021/file/01632f7b7a127233fa1188bd6c2e42e1-Paper.pdf
Neal R. M., 1994, J. R. Stat. Soc. A, 56, 41
Newton M. A., Raftery A. E., 1994, J. R. Stat. Soc. B, 56, 3
Papamakarios G., Murray I., 2016, Proc. 30th International Conference on Neural Information Processing Systems. NIPS'16. Curran Associates, Inc., NY, USA, p. 1036
Papamakarios G., Pavlakou T., Murray I., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., eds, Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., NY, USA. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf
Papamakarios G., Sterratt D., Murray I., 2019, in Chaudhuri K., Sugiyama M., eds, Proc. Mach. Learn. Res. Vol. 89, Proc. 22nd International Conference on Artificial Intelligence and Statistics. PMLR, p. 837. Available at: https://proceedings.mlr.press/v89/papamakarios19a.html
Papamakarios G., Nalisnick E., Rezende D. J., Mohamed S., Lakshminarayanan B., 2021, J. Mach. Learn. Res., 22, 1
Robert C. P., Wraith D., 2009, AIP Conf. Proc. Vol. 1193, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 251
Skilling J., 2006, Bayesian Anal., 1, 833
Taracchini A. et al., 2014, Phys. Rev. D, 89, 061502
Tejero-Cantero A., Boelts J., Deistler M., Lueckmann J.-M., Durkan C., Gonçalves P. J., Greenberg D. S., Macke J. H., 2020, J. Open Source Softw., 5, 2505
Trotta R., 2007, MNRAS, 378, 72

Trotta R., 2008, Contemp. Phys., 49, 71
van Haasteren R., 2014, Gravitational Wave Detection and Data Analysis for Pulsar Timing Arrays. Springer, Berlin, p. 99

SUPPORTING INFORMATION

Supplementary data are available at *RASTAI* online.

Harmonic_SBI_final_submission.zip

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.