

Estimating Marginal Likelihoods in Likelihood-Free Inference via Neural Density Estimation

Paul Bastide¹, Arnaud Estoup², Jean-Michel Marin³, and Julien Stoehr^{4, 5}

¹Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

²CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Université Montpellier, Montpellier, France

³IMAG, Université de Montpellier, CNRS, 34090 Montpellier, France

⁴Université Paris-Dauphine, Université PSL, CNRS, CEREMADE, 75016 Paris, France

⁵Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France

July, 2025

Abstract

The marginal likelihood, or evidence, plays a central role in Bayesian model selection, yet remains notoriously challenging to compute in likelihood-free settings. While Simulation-Based Inference (SBI) techniques such as Sequential Neural Likelihood Estimation (SNLE) offer powerful tools to approximate posteriors using neural density estimators, they typically do not provide estimates of the evidence. In this technical report presented at BayesComp 2025, we present a simple and general methodology to estimate the marginal likelihood using the output of SNLE.

1 Introduction

1.1 Context

Bayesian inference provides a principled framework for learning from data by combining prior information with observed evidence. A central quantity in this framework is the *marginal likelihood* or *evidence*, defined as the model likelihood integrated over the prior distribution, namely

$$p(\mathbf{x}) = \int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

The latter serves as the normalizing constant in Bayes' theorem and underpins model comparison. However, computing this integral is often intractable, particularly in complex models where the likelihood $f(\mathbf{x}|\boldsymbol{\theta})$ is unavailable in closed form or prohibitively expensive to evaluate. Simulation-Based Inference (SBI), also known as likelihood-free inference, addresses this challenge by circumventing the need for an explicit likelihood function. In the typical setting, we assume a generative model $f(\mathbf{x} | \boldsymbol{\theta})$ from which simulations can be drawn for any parameter value $\boldsymbol{\theta} \in \Theta$, even though the likelihood function itself is not pointwise evaluable. A proper prior $\pi(\boldsymbol{\theta})$ is specified, and the goal becomes to infer the posterior $p(\boldsymbol{\theta} | \mathbf{x}^*)$ from simulations alone. One of the earliest and most influential families of methods in this space is Approximate Bayesian Computation (ABC), which proceeds by simulating pseudo-datasets and comparing them to the observed dataset via summary statistics and a tolerance criterion. Since the foundational work of [Tavare et al. \[1997\]](#) and [Pritchard et al. \[1999\]](#), ABC has evolved significantly, often incorporating machine learning techniques to improve efficiency and robustness [[Pudlo et al., 2016](#), [Sheehan and Song, 2016](#)]. More recently, the SBI landscape has been transformed by the introduction of neural network-based techniques, which leverage modern density estimation and amortized inference tools. These methods are typically categorized by the quantity they aim to approximate:

- **NPE** and **SNPE** (Sequential Neural Posterior Estimation) targets the posterior directly;
- **NLE** and **SNLE** (Sequential Neural Likelihood Estimation) learns a surrogate for the likelihood;
- **NRE** and **SNRE** (Sequential Neural Ratio Estimation) focuses on likelihood ratios.

Notable contributions in this space include Papamakarios and Murray [2016], Papamakarios et al. [2019], Greenberg et al. [2019], Hermans et al. [2020], Cranmer et al. [2020], and a comprehensive benchmarking study by Lueckmann et al. [2021] provides a detailed comparison of these approaches.

While these neural SBI methods have significantly improved the quality of posterior inference, they are typically not designed to provide estimates of the marginal likelihood (with the notable recent exception of Spurio Mancini et al. 2023, see below). As a result, Bayesian model comparison, a difficult task in the likelihood-free setting, is dominated by ABC methodologies [Grelaud et al., 2009, Pudlo et al., 2016, Marin et al., 2018]. In this work, we focus on addressing this gap by proposing a family of methods that leverage the output of SNLE to estimate the marginal likelihood

$$C = \int f(\mathbf{x}^* | \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \quad (1)$$

for some observed data \mathbf{x}^* , despite the intractability of $f(\mathbf{x}^* | \boldsymbol{\theta})$.

1.2 Sequential Neural Likelihood Estimation

Sequential Neural Likelihood Estimation (SNLE) is a powerful approach within SBI that approximates the likelihood function $f(\mathbf{x} | \boldsymbol{\theta})$ using a neural density estimator. This approximation is constructed iteratively over L rounds of adaptive simulation, refining the learned likelihood by focusing simulations in regions of high posterior probability (see Algorithm 1). In this work, we focus specifically on density estimators based on normalizing flows (NF) [Papamakarios et al., 2021].

After completing all rounds, samples from the final approximate posterior are obtained via MCMC targeting $\hat{\pi}^{(L)}(\boldsymbol{\theta} | \mathbf{x}^*)$. SNLE has shown to match or surpass other SBI methods in terms of inference quality, while often requiring significantly fewer simulations. That said, SNLE requires careful tuning of its components—architecture of the density estimator, learning rate schedules, and sampling strategy for MCMC. It also inherits the sensitivity to model misspecification common in simulation-based approaches.

SNLE is a prominent method within simulation-based inference (SBI) that focuses on iteratively refining a surrogate likelihood using rounds of targeted simulations. Unlike fully amortized approaches such as Neural Posterior Estimation (NPE) or Neural Likelihood Estimation (NLE), which train a single model to generalize across all possible observations, SNLE concentrates computational effort on the region of parameter space relevant to a specific observed dataset \mathbf{x}^* . This sequential adaptation enables SNLE to achieve higher accuracy with fewer simulations, making it especially effective in settings where simulation is expensive.

Amortized inference methods like those proposed by Papamakarios and Murray [2016] and further developed by Cranmer et al. [2020] aim to build models that generalize across multiple inference tasks. While this amortization offers efficiency for repeated queries, it may underperform in single-instance inference due to limited expressivity or training resource constraints. In contrast, SNLE trades amortization for focus: each round adapts the simulator budget to the posterior’s high-probability region, leading to more precise inference at the cost of reuse across datasets.

While SNLE was originally introduced as a method for posterior inference in likelihood-free settings, its outputs contain richer information that can be exploited beyond posterior approximation. In particular, SNLE produces a surrogate likelihood function $q^{(L)}(\mathbf{x} | \boldsymbol{\theta})$ and uses it in combination with the prior $\pi(\boldsymbol{\theta})$ to construct an approximate posterior distribution.

In this work, we show that these components can also be used to estimate the marginal likelihood C . Indeed, although the true likelihood cannot be evaluated, SNLE provides a trained surrogate $q^{(L)}(\mathbf{x}^* | \boldsymbol{\theta})$ that approximates it in regions of high posterior mass. Our goal is to develop practical strategies for estimating the marginal likelihood C using only this surrogate likelihood and the posterior samples generated during SNLE’s sequential training procedure. These strategies enable Bayesian model comparison in simulation-based contexts, where the marginal likelihood is otherwise inaccessible.

Algorithm 1: SNLE [Papamakarios et al., 2019]

Input: observed dataset \mathbf{x}^* , prior distribution $\pi(\cdot)$, simulator $f(\cdot | \boldsymbol{\theta})$, number of iterations L , number of simulations per iteration N

Output: Trained posterior estimator $\hat{\pi}^{(L)}(\cdot | \mathbf{x}^*)$

Set initial proposal $\hat{\pi}^{(0)}(\cdot | \mathbf{x}^*) = \pi(\boldsymbol{\theta})$ and training dataset $\mathcal{D} = \emptyset$

for $\ell = 1$ to L **do**

- for** $i = 1$ to N **do**
- Sample $\boldsymbol{\theta}_i^{(\ell)} \sim \hat{\pi}^{(\ell-1)}(\cdot | \mathbf{x}^*)$
- Sample $\mathbf{x}_i^{(\ell)} \sim f(\cdot | \boldsymbol{\theta}_i^{(\ell)})$
- Add $(\boldsymbol{\theta}_i^{(\ell)}, \mathbf{x}_i^{(\ell)})$ to \mathcal{D}
- end**
- Train $q^{(\ell)}(\mathbf{x} | \boldsymbol{\theta})$ on dataset \mathcal{D} using maximum likelihood
- Set $\hat{\pi}^{(\ell)}(\cdot | \mathbf{x}^*) \propto q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta})\pi(\boldsymbol{\theta})$
- end**

return $\hat{\pi}^{(L)}(\cdot | \mathbf{x}^*)$

2 The SIS-SNLE formulation

We introduce here a Sequential Importance Sampling (SIS) [Owen, 2013, Robert and Casella, 2004] technique, that can leverage the iterative nature of SNLE to progressively build an estimate of the marginal likelihood.

Let $q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta})$ denote the neural likelihood approximation obtained after round ℓ of SNLE. We can then define a surrogate for the evidence associated with this intermediate approximation at round ℓ as

$$C_\ell = \int q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

The final estimate C_L after L rounds is expected to closely approximate the true marginal likelihood (1).

We now introduce a SIS estimator for C_L , based on evaluating the ratio $C_\ell/C_{\ell-1}$ across rounds. Assuming the prior distribution is absolutely continuous with respect to the SNLE surrogate posterior $\hat{\pi}^{(\ell-1)}(\boldsymbol{\theta} | \mathbf{x}^*)$, a simple change of measure leads to

$$\begin{aligned} C_\ell &= \int q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \int q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta})}{\hat{\pi}^{(\ell-1)}(\boldsymbol{\theta} | \mathbf{x}^*)} \hat{\pi}^{(\ell-1)}(\boldsymbol{\theta} | \mathbf{x}^*)d\boldsymbol{\theta}, \end{aligned}$$

where the posterior from the previous round satisfies

$$\hat{\pi}^{(\ell-1)}(\boldsymbol{\theta} | \mathbf{x}^*) = \frac{\pi(\boldsymbol{\theta})q^{(\ell-1)}(\mathbf{x}^* | \boldsymbol{\theta})}{C_{\ell-1}}.$$

It follows that

$$R_\ell = \frac{C_\ell}{C_{\ell-1}} = \int \frac{q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta})}{q^{(\ell-1)}(\mathbf{x}^* | \boldsymbol{\theta})} \hat{\pi}^{(\ell-1)}(\boldsymbol{\theta} | \mathbf{x}^*) d\boldsymbol{\theta}.$$

This ratio can be estimated by Monte Carlo using the samples $(\boldsymbol{\theta}_i^{(\ell-1)})_{1 \leq i \leq N}$ from $\hat{\pi}^{(\ell-1)}(\boldsymbol{\theta} | \mathbf{x}^*)$ that were generated at iteration $\ell - 1$ of Algorithm 1, namely

$$\widehat{R}_\ell = \frac{1}{N} \sum_{i=1}^N \frac{q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta}_i^{(\ell-1)})}{q^{(\ell-1)}(\mathbf{x}^* | \boldsymbol{\theta}_i^{(\ell-1)})}.$$

Letting $C_0 = 1$ and $q^{(0)}(\cdot | \boldsymbol{\theta}) \equiv 1$, we estimate the final evidence as the product:

$$\widehat{C}_L = \prod_{\ell=1}^L \widehat{R}_\ell = \prod_{\ell=1}^L \left(\frac{1}{N} \sum_{i=1}^N \frac{q^{(\ell)}(\mathbf{x}^* | \boldsymbol{\theta}_i^{(\ell-1)})}{q^{(\ell-1)}(\mathbf{x}^* | \boldsymbol{\theta}_i^{(\ell-1)})} \right). \quad (2)$$

This estimator can be computed directly from the sequence of density approximations and posterior samples produced by Algorithm 1. Conceptually, it draws inspiration from classical methods that estimate partition functions or marginal likelihoods by exploiting a sequence of intermediate distributions.

- The **Steppingstone Sampling (SS)** estimator [Xie et al., 2011] constructs a geometric path between the prior and posterior by introducing a series of tempered (power) posteriors. Each intermediate distribution bridges the gap between prior and posterior, allowing for stable importance weight updates across steps. The marginal likelihood is then obtained as a product of ratios of normalizing constants between consecutive steps, similar in spirit to the recursive product formulation in SIS-SNLE.
- **Sequential Monte Carlo (SMC)** samplers [Moral et al., 2006] also construct a sequence of intermediate distributions and estimate normalizing constants using weighted particle approximations. These methods rely on importance sampling and resampling steps to control the variance and particle degeneracy across the sequence.

In our setting, the rounds of the SNLE algorithm naturally yield a sequence of increasingly accurate posterior approximations. As the neural likelihood surrogates are refined, the corresponding posterior approximations become more concentrated around high-probability regions. This structure mirrors the gradual transition seen in both SS and SMC approaches, making the Sequential Importance Sampling strategy a natural and computationally efficient tool for marginal likelihood estimation within the SNLE framework.

3 The IS-SNLE method

The SIS-SNLE estimator (2) described in the previous section leverages the intermediate approximations produced during SNLE, but it does not use MCMC samples from the final posterior approximation $\widehat{\pi}^{(L)}(\boldsymbol{\theta} | \mathbf{x}^*)$ that is used for parameter estimation. This observation motivates a complementary strategy based on standard Importance Sampling (IS) [Owen, 2013, Robert and Casella, 2004] using the final posterior approximation. Let $(\boldsymbol{\theta}_i^{(L)})_{1 \leq i \leq N}$ denote the MCMC sample from $\widehat{\pi}^{(L)}(\boldsymbol{\theta} | \mathbf{x}^*)$. The IS-SNLE strategy proceeds in three steps:

- Train a neural density estimator $h(\boldsymbol{\theta})$ on the MCMC sample $(\boldsymbol{\theta}_i^{(L)})_{1 \leq i \leq N}$ to approximate the marginal posterior $\widehat{\pi}^{(L)}(\boldsymbol{\theta} | \mathbf{x}^*)$
- Generate a new, independent and identically distributed (i.i.d.) sample $(\boldsymbol{\theta}_j^{IS})_{1 \leq j \leq N_{IS}}$
- Estimate the marginal likelihood C_L using the importance sampling estimator

$$\widehat{C}_L = \frac{1}{N_{IS}} \sum_{j=1}^{N_{IS}} q^{(L)}(\mathbf{x}^* | \boldsymbol{\theta}_j^{IS}) \frac{\pi(\boldsymbol{\theta}_j^{IS})}{h(\boldsymbol{\theta}_j^{IS})}. \quad (3)$$

Importance sampling can use any proper distribution as the sampling distribution, but would be optimal if one were able to use the posterior distribution. As it is unknown, we propose here to replace it by a neural estimate. This method hence fully exploits two crucial properties of neural density estimators using normalizing flows [Papamakarios et al., 2021]: (i) the estimation $h(\boldsymbol{\theta})$ is a proper density on the space of parameters Θ , making it suitable for IS, and (ii) the very construction of $h(\boldsymbol{\theta})$ allows efficient sampling from this distribution, so that we can generate the N_{IS} independent draws easily and at a low computational cost.

This estimator requires only the final surrogate likelihood approximation $q^{(L)}(\mathbf{x}^* | \boldsymbol{\theta})$, the prior density $\pi(\boldsymbol{\theta})$, and a tractable density estimate $h(\boldsymbol{\theta})$ derived from the MCMC output. Crucially, the method is not specific to simulation-based inference and could be applied in any Bayesian setting where a posterior sample is available and a surrogate likelihood is accessible.

4 The HM-SNLE method

In addition to importance sampling-based approaches, one may consider the classical Harmonic Mean (HM) estimator for marginal likelihood [Newton and Raftery, 1994]. Following early work by McEwen et al. [2021], Spurio Mancini et al. [2023] propose a retargeted version of this estimator tailored to simulation-based inference, which leverages the surrogate likelihood learned by SNLE. Remember that $q^{(L)}(\mathbf{x}^* \mid \boldsymbol{\theta})$ denote the neural likelihood approximation after L rounds and that $\pi^{(L)}(\boldsymbol{\theta} \mid \mathbf{x}^*)$ denote the corresponding posterior. The marginal likelihood can then be expressed as:

$$C_L^{-1} = \int \left\{ \frac{\psi(\boldsymbol{\theta})}{q^{(L)}(\mathbf{x}^* \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})} \right\} \pi^{(L)}(\boldsymbol{\theta} \mid \mathbf{x}^*) d\boldsymbol{\theta}$$

where $\psi(\boldsymbol{\theta})$ is any normalized density used to stabilize the estimator. This design helps mitigate the notorious instability of the harmonic mean estimator, which otherwise suffers from infinite variance when the denominator becomes too small in the tails.

Just like the IS case, the optimal choice for $\psi(\boldsymbol{\theta})$ would be the unknown posterior distribution. Spurio Mancini et al. [2023] hence propose to estimate $\psi(\boldsymbol{\theta})$ as a normalizing flow. To avoid the double use of the data, the final MCMC sample from the posterior $\hat{\pi}^{(L)}(\boldsymbol{\theta} \mid \mathbf{x}^*)$ is split between an evaluating set $(\boldsymbol{\theta}_i^{(L)})_{1 \leq i \leq N_{\text{eval}}}$ and a learning set $(\boldsymbol{\theta}_i^{(L)})_{N_{\text{eval}}+1 \leq i \leq N}$, with $1 \leq N_{\text{learn}} \leq N$ and $N_{\text{eval}} = N - N_{\text{learn}}$. The learning set is used to fit a NF to estimate $\psi(\boldsymbol{\theta})$, and the evaluating set to fitted to a subset of the posterior samples generated via MCMC from $\hat{\pi}^{(L)}(\boldsymbol{\theta} \mid \mathbf{x}^*)$. The final estimator is then

$$\widehat{C}_L^{-1} = \frac{1}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} \left\{ \frac{\psi(\boldsymbol{\theta}_i^{(L)})}{q^{(L)}(\mathbf{x}^* \mid \boldsymbol{\theta}_i^{(L)})\pi(\boldsymbol{\theta}_i^{(L)})} \right\}. \quad (4)$$

This estimator is simple to compute, leverages existing posterior samples, and avoids additional simulator calls. However, it remains sensitive to the tail behavior of the reweighting terms, and careful design of ψ is essential—particularly in high-dimensional spaces where tail mismatch can still lead to unstable estimates.

5 Concentrated or dilated normalizing flows

It is well known in the literature that the HM estimator benefits from a proposal $\psi(\boldsymbol{\theta})$ with lighter tails than the posterior. To enforce such a behavior, Polanska et al. [2024] propose to use the very structure of the NF, by manually changing the variance of the base distribution of the flow *after it has been trained on the posterior sample*. They apply a so-called multiplicative “temperature” T on the variance of the base distribution of the flow, with $0 < T \leq 1$. The following transformation layers of the flow remain unchanged. As the estimator (4) is valid for any proper distribution, $\psi_T(\boldsymbol{\theta})$ can then readily be used, in the formula, and provides for a more concentrated instrumental distribution.

A similar trick can be used for our IS technique, with the crucial difference that the proposal distribution in this case must have heavy tails compared to the posterior distribution. We hence propose to simply multiply the variance of the base distribution of the learned flow by a temperature T , but with $T \geq 1$, so as to dilute the proposal distribution, instead of concentrating it. The resulting NF $h_T(\boldsymbol{\theta})$ is still a proper distribution that is easy to sample from by construction.

6 Numerical experiments

We consider the simple Gaussian toy example described in Spurio Mancini et al. [2023]. It is defined by the generative model

$$\mathbf{x} = (x_1, \dots, x_d) \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, I_d), \quad \text{with } \boldsymbol{\theta} = (\theta_1, \dots, \theta_d) \sim \mathcal{U}[-2, 2]^d,$$

and we fix the observed dataset to $\mathbf{x}^* = (0, \dots, 0)$. In this setting, the marginal likelihood admits a closed-form expression

$$C = \frac{1}{4^d (2\pi)^{d/2}} \int_{-2}^2 \cdots \int_{-2}^2 \exp \left(-\frac{1}{2} \sum_{i=1}^d \theta_i^2 \right) d\theta_1 \cdots d\theta_d = \frac{[\text{erf}(\sqrt{2})]^d}{4^d},$$

where erf denotes the Gauss error function [Spurio Mancini et al., 2023]. The SNLE configuration is as follow

- $L = 5$ rounds, each with $N = 1,000$ simulations.
- Likelihood model: Masked Autoregressive Flow (MAF) [Papamakarios et al., 2017] with 5 transformations, each with 2 hidden layers of 64 neurons.
- Early stopping after 20 epochs without validation improvement (10% of data used for validation).
- Posterior inference via slice sampling using 20 parallel chains, producing 1,000 approximate posterior samples.

The SNLE procedure is implemented using `zuko v1.4.0` [Rozet et al., 2022] through the `sbi v0.24` interface Boelts et al. [2025]. Note that, compared with Spurio Mancini et al. [2023] that used a SNLE with 10 rounds of 10,000 simulations each, our inference setup is much lighter, which illustrates the better efficiency of IS and SIS methods compared to HM.

For the Importance Sampling (IS) estimator, we generate $N = 1,000$ samples from $\hat{\pi}^{(L)}(\boldsymbol{\theta} | \mathbf{x}^*)$ using the same slice sampler with 20 parallel chains, and train on this sample a MAF with 3 transformations, each with two hidden layers of 32 neurons. We then vary the temperature from 1.0 (no tempering) to 2, taking 1.25 as a default, and generate 1,000 new independent samples from the learned flow $h_T(\boldsymbol{\theta})$.

For the Harmonic Mean (HM) estimator, we follow the setup of Spurio Mancini et al. [2023], Polanska et al. [2024]. We first generate $N' = 2,000$ samples from $\hat{\pi}^{(L)}(\boldsymbol{\theta} | \mathbf{x}^*)$ using the same slice sampler with 20 parallel chains, and split this sample between a 1,000 training set and a 1,000 evaluating set. We generated twice as many samples from the posterior compared to IS, so that the same number of 1,000 samples can be used for training and evaluation in both cases. The normalized density $\psi(\boldsymbol{\theta})$ is then modeled using a Real-valued Non-Volume Preserving (RealNVP) [Dinh et al., 2017] NF, and we apply a temperature T varying between 0.5 and 1.0, with a default to 0.8. This estimator is implemented using the `harmonic v1.2.3` package [McEwen et al., 2021].

As the inference method is stochastic, we re-ran all the algorithms 25 times to account for the variability of the results on a fixed dataset \mathbf{x}^* .

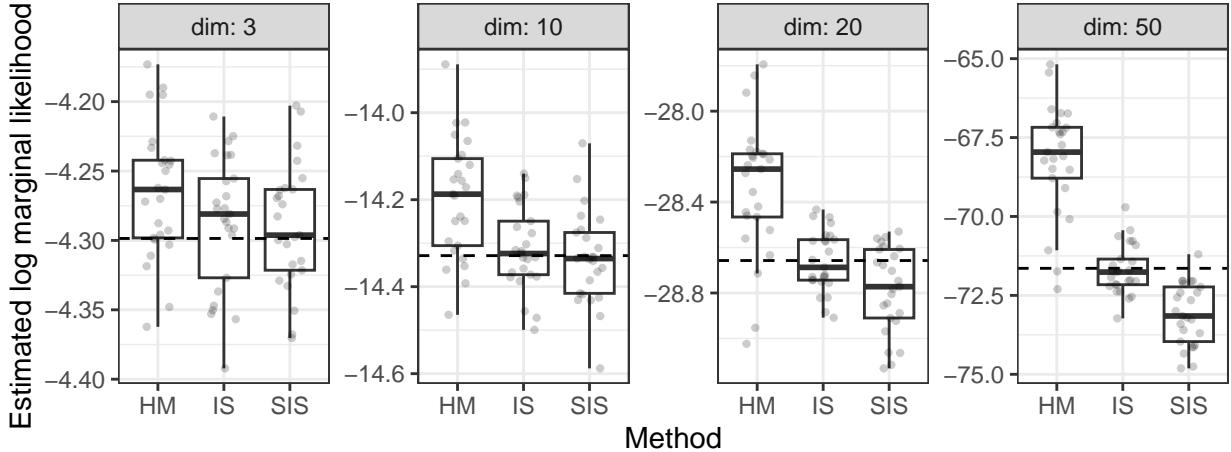


Figure 1: Estimated log-marginal likelihoods with HM, IS and SIS. Each panel corresponds to a dimension d , and the dashed line corresponds to the true log-marginal likelihoods. For each method, 25 independent estimates are shown, highlighting both accuracy and variability. We used a default temperature of, respectively, 0.8 for HM and $1/0.8 = 1.25$ for IS.

Figure 1 illustrates that the HM estimator performed poorly as the dimensionality increased, exhibiting high variance and bias. The IS estimator consistently outperformed the SIS approach in higher-dimensional

settings. Figure 2 and 3 show that the HM estimator was quite sensitive to the temperature parameter, while IS seemed to be more robust to variations of it.

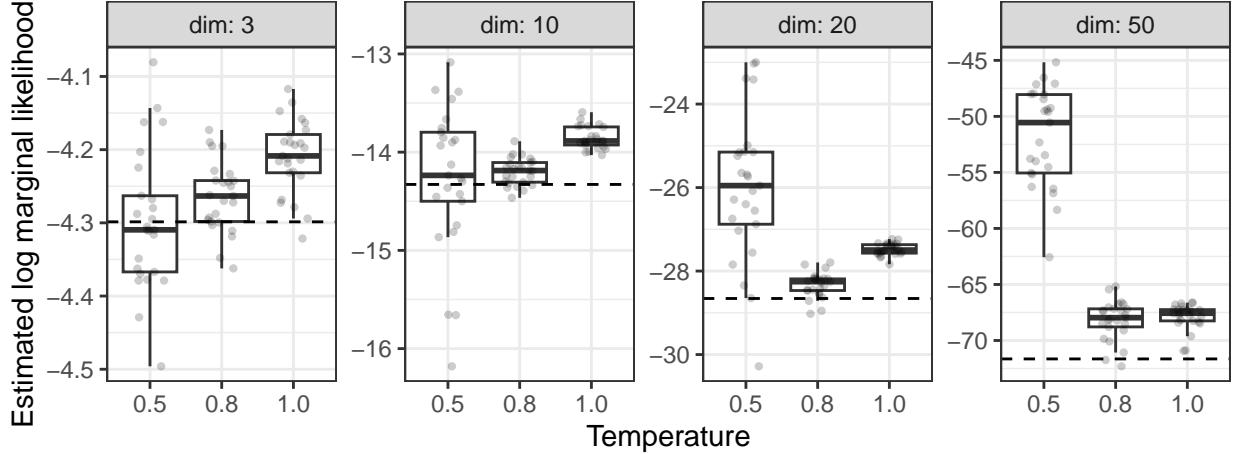


Figure 2: Estimated log-marginal likelihoods with HM and varying temperature. Each panel corresponds to a dimension d , and the dashed line corresponds to the true log-marginal likelihoods. For each method, 25 independent estimates are shown.

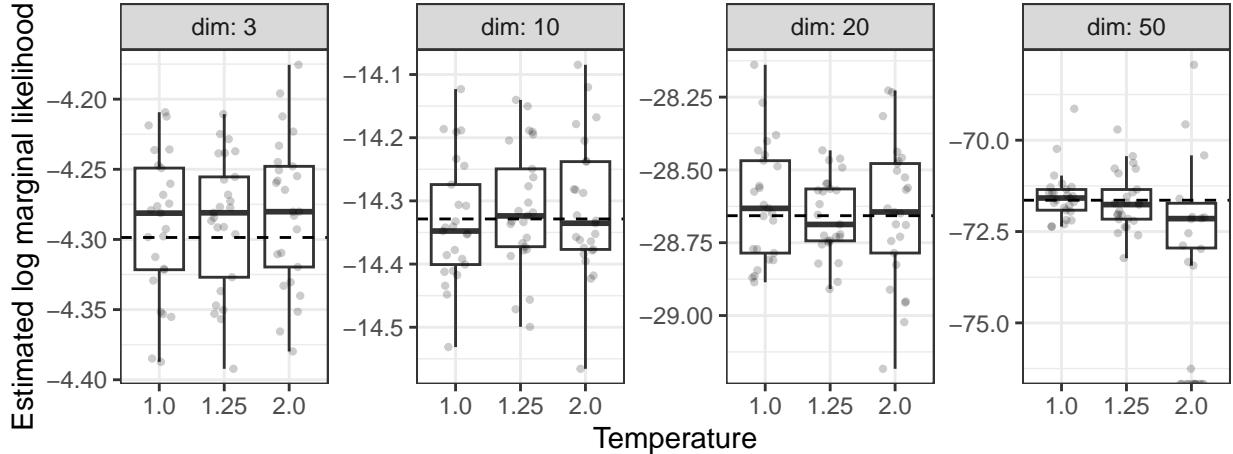


Figure 3: Estimated log-marginal likelihoods with IS and varying temperature. Each panel corresponds to a dimension d , and the dashed line corresponds to the true log-marginal likelihoods. For each method, 25 independent estimates are shown.

7 Discussion

We introduced a general method to estimate the marginal likelihood in likelihood-free models using SNLE outputs. Our approach is simple, generic, and computationally efficient, leveraging the structure of SNLE to estimate the evidence via importance sampling. Challenges include potential variance due to support mismatch between the posterior and the proposal distribution, and the propagation of density estimation

errors. Nonetheless, this provides a foundation for applying Bayesian model selection tools in the SBI setting. In addition to more extensive testing in more complex or realistic settings, future research will investigate variance-reducing techniques, alternative proposal constructions, and robustness diagnostics. Extending this methodology to SNPE and SNRE is also a promising direction.

References

- J. Boelts, M. Deistler, M. Glöckler, Á. Tejero-Cantero, J.-M. Lueckmann, G. Moss, P. Steinbach, T. Moreau, F. Muratore, J. Linhart, C. Durkan, J. Vetter, B. K. Miller, M. Herold, A. Ziaeemehr, M. Pals, T. Gruner, S. Bischoff, N. Kruglova, R. Gao, J. K. Lappalainen, B. Mucsányi, F. Pei, A. Schulz, Z. Stefanidi, P. Rodrigues, C. Schröder, F. Abu Zaid, J. Beck, J. Kapoor, D. S. Greenberg, P. J. Gonçalves, and J. H. Macke. sbi reloaded: A toolkit for simulation-based inference workflows. *The Journal of Open Source Software*, 10(108):7754, 2025. doi: 10.21105/joss.07754. URL <https://doi.org/10.21105/joss.07754>.
- Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. February 2017. URL <https://openreview.net/forum?id=HkpbnH9lx>.
- David Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference. In *ICML*, 2019.
- A. Grelaud, J.-M. Marin, C. P. Robert, F. Rodolphe, and F. Taly. Likelihood-free methods for model choice in gibbs random fields. *Bayesian Analysis*, 3(2):427–442, 2009. doi: 10.1214/09-BA412.
- Joeri Hermans, Vincent Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate likelihood ratios. In *ICML*, 2020.
- Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Gonçalves, and Jakob H. Macke. Benchmarking simulation-based inference. In *AISTATS*, 2021.
- J.-M. Marin, P. Pudlo, A. Estoup, and C.P. Robert. Likelihood-free model choice. In S. A. Sisson, Y. Fan, and M. Beaumont, editors, *Handbook of Approximate Bayesian Computation*, chapter 6. Chapman & Hall/CRC, 2018.
- 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 2021. doi: 10.48550/ARXIV.2111.12720. URL <http://arxiv.org/abs/2111.12720>. arXiv:2111.12720 [stat].
- Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006. doi: 10.1111/j.1467-9868.2006.00553.x.
- Michael A. Newton and Adrian E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.
- Art B. Owen. Monte carlo theory, methods and examples. 2013. Book manuscript, available at <https://statweb.stanford.edu/~owen/mc/>.
- George Papamakarios and Iain Murray. Fast ϵ -free inference of simulation models with bayesian conditional density estimation. *NeurIPS*, 2016.
- George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *AISTATS*, 2019.

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. URL <http://jmlr.org/papers/v22/19-1028.html>.

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. 2024. doi: 10.48550/ARXIV.2405.05969. URL <https://arxiv.org/pdf/2405.05969.pdf>.

Jonathan K. Pritchard, M.T. Siestad, A. Perez-Lezaun, and M.W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P. Robert. Reliable abc model choice via random forests. *Bioinformatics*, 32(6):859–866, 2016.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004. ISBN 9780387212395.

François Rozet et al. Zuko: Normalizing flows in pytorch, 2022. URL <https://pypi.org/project/zuko>.

Sara Sheehan and Yun S. Song. Deep learning for population genetic inference. *PLOS Computational Biology*, 12(3):e1004845, 2016.

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, 2023. doi: 10.1093/rasti/rzad051.

Simon Tavaré, David J. Balding, Robert C. Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

Wangang Xie, Paul O. Lewis, Yu Fan, Lynn Kuo, and Ming-Hui Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160, 2011. doi: 10.1093/sysbio/syq085.