

Marginal likelihood computation for model selection and hypothesis testing: an extensive review

F. Llorente^{*}, L. Martino^{**}, D. Delgado^{*}, J. Lopez-Santiago^{*}

^{*} Universidad Carlos III de Madrid, Leganés (Spain).

^{**} Universidad Rey Juan Carlos, Fuenlabrada (Spain).

Abstract

This is an up-to-date introduction to, and overview of, marginal likelihood computation for model selection and hypothesis testing. Computing normalizing constants of probability models (or ratio of constants) is a fundamental issue in many applications in statistics, applied mathematics, signal processing and machine learning. This article provides a comprehensive study of the state-of-the-art of the topic. We highlight limitations, benefits, connections and differences among the different techniques. Problems and possible solutions with the use of improper priors are also described. Some of the most relevant methodologies are compared through theoretical comparisons and numerical experiments.

Keywords: Marginal likelihood, Bayesian evidence, numerical integration, model selection, hypothesis testing, quadrature rules, double-intractable posteriors, partition functions

1 Introduction

Marginal likelihood (a.k.a., Bayesian evidence) and Bayes factors are the core of the Bayesian theory for testing hypotheses and model selection [1, 2]. More generally, the computation of normalizing constants or ratios of normalizing constants has played an important role in statistical physics and numerical analysis [3]. In the Bayesian setting, the approximation of normalizing constants is also required in the study of the so-called double intractable posteriors [4].

Several methods have been proposed for approximating the marginal likelihood and normalizing constants in the last decades. Most of these techniques have been originally introduced in the field of statistical mechanics. Indeed, the marginal likelihood is the analogous of a central quantity in statistical physics known as the *partition function* which is also closely related to another important quantity often called *free-energy*. The relationship between statistical physics and Bayesian inference has been remarked in different works [5, 6].

The model selection problem has been also addressed from different points of view. Several criteria have been proposed to deal with the trade-off between the goodness-of-fit of the model and its simplicity. For instance, the Akaike information criterion (AIC) or the focused information criterion (FIC) are two examples of these approaches [7, 8]. The Bayesian-Schwarz information

criterion (BIC) is related to the marginal likelihood approximation, as discussed in Section 3. The deviance information criterion (DIC) is a generalization of the AIC, which is often used in Bayesian inference [9, 10]. It is particularly useful for hierarchical models and it can be approximately computed when the outputs of a Markov Chain Monte Carlo (MCMC) algorithm are given. However, DIC is not directly related to the Bayesian evidence [11]. Another different approach, also based on information theory, is the so-called minimum description length principle (MDL) [12]. MDL was originally derived for data compression, and then was applied to model selection and hypothesis testing. Roughly speaking, MDL considers that the best explanation for a given set of data is provided by the *shortest description* of that data [12].

In the Bayesian framework, there are two main classes of sampling algorithms. The first one consists in approximating the marginal likelihood of different models or the ratio of two marginal likelihoods. In this work, we focus on this first approach. The second sampling approach extends the posterior space including a discrete indicator variable m , denoting the m -th model [13, 14]. For instance, in the well-known **reversible jump MCMC** [14], a Markov chain is generated in this extended space, allowing jumps between models with possibly different dimensions. However, generally, these methods are difficult to tune and the mixing of the chain can be poor [15]. For further details, see also the interesting works [16, 17, 18]. The average number of MCMC iterations when the chain jumps or stays into the m -th model is proportional to the marginal likelihood of the corresponding model.

In this work, we provide an extensive review of computational techniques for the marginal likelihood computation. The main contribution is to present jointly numerous computational schemes (introduced independently in the literature) with a detailed description under the same notation, highlighting their differences, relationships, limitations and strengths. Most of them are based on the importance sampling (IS) approach and several of them are combination the MCMC and IS schemes. It is also important to remark that parts of the presented material are also novel, i.e., no contained in previous works. We have widely studied, analyzed and jointly described with a unique notation and classification, the methodologies presented in a vast literature from 1990s to the recent proposed algorithms (see Table 24). We also discuss issues and solutions when improper priors are employed. Therefore, this survey provides an ample covering of the literature, where we highlight important details and comparisons in order to facilitate the understanding of the interested readers and practitioners.

The problem statement and the main notation are introduced in the Section 2.1. Relevant considerations regarding the marginal likelihood and other model selection strategies are given in Section 2.2 and Section 7. Specifically, a description of how the marginal likelihood handles the model fit and the model complexity is provided in Section 2.2. The dependence on the prior selection and the possible choice of an improper prior are discussed in Section 7. The different techniques have been classified in four main families, as shown in Section 2.3. Sections 3, 4, 5, 6 are devoted to the detailed description of the computational schemes for approximating the Bayesian evidence. Section 8 contains some numerical experiments. In Section 9, we conclude with a final summary and discussion. We provide also theoretical analyses of some of the experiments and other comparisons in the Supplementary Material.

Contents

1	Introduction	1
2	Problem statement and preliminary discussions	4
2.1	Framework and notation	4
2.2	Model fit and model complexity	6
2.3	A general overview of the computational methods	8
3	Methods based on deterministic approximations and density estimation	9
3.1	Laplace’s method	9
3.2	Bayesian-Schwarz information criterion (BIC)	10
3.3	Kernel density estimation (KDE)	11
3.4	Chib’s method	11
3.5	Interpolative approaches	12
4	Techniques based on IS	13
4.1	Techniques using draws from one proposal density	15
4.2	Techniques using draws from two proposal densities	21
4.3	IS based on multiple proposal densities	27
5	Advanced schemes combining MCMC and IS	36
5.1	MCMC-within-IS: weighted samples after MCMC iterations	36
5.2	Weighted samples after MCMC and resampling steps	39
5.3	IS-within-MCMC: Estimation based on Multiple Try MCMC schemes	44
5.4	IS-after-MCMC: Layered Adaptive Importance Sampling (LAIS)	45
6	Vertical likelihood representations	48
6.1	Lebesgue representations of the marginal likelihood	49
6.2	Nested Sampling	53
7	On the marginal likelihood approach and other strategies	57
7.1	Dependence on the prior and related discussion	58
7.2	Bayes factors with improper priors	59
7.3	Marginal likelihood as a prior predictive approach	62
7.4	Other ways of model selection: the posterior predictive approach	62
8	Numerical comparisons	64
8.1	First experiment	64
8.2	Second experiment: Gaussian likelihood and uniform prior	68
8.3	Third experiment: posterior as mixture of two components	71
8.4	Experiment with biochemical oxygen demand data	73
8.5	Experiment with COVID-19 data	75
9	Final discussion	77

2 Problem statement and preliminary discussions

2.1 Framework and notation

In many applications, the goal is to make inference about a variable of interest, $\boldsymbol{\theta} = \theta_{1:D_\theta} = [\theta_1, \theta_2, \dots, \theta_{D_\theta}] \in \Theta \subseteq \mathbb{R}^{D_\theta}$, where $\theta_d \in \mathbb{R}$ for all $d = 1, \dots, D_\theta$, given a set of observed measurements, $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$. In the Bayesian framework, one complete model \mathcal{M} is formed by a likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$ and a prior probability density function (pdf) $g(\boldsymbol{\theta}|\mathcal{M})$. All the statistical information is summarized by the posterior pdf, i.e.,

$$P(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})g(\boldsymbol{\theta}|\mathcal{M})}{p(\mathbf{y}|\mathcal{M})}, \tag{1}$$

where

$$Z = p(\mathbf{y}|\mathcal{M}) = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})g(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}, \tag{2}$$

is the so-called marginal likelihood, a.k.a., Bayesian evidence. This quantity is important for model selection purpose, as we show below. However, usually $Z = p(\mathbf{y}|\mathcal{M})$ is unknown and difficult to approximate, so that in many cases we are only able to evaluate the unnormalized target function,

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})g(\boldsymbol{\theta}|\mathcal{M}). \tag{3}$$

Note that $P(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}) \propto \pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M})$ [1, 2]. For the sake of simplicity, hereafter we use the simplified notation $P(\boldsymbol{\theta}|\mathbf{y})$ and $\pi(\boldsymbol{\theta}|\mathbf{y})$. Thus, note that

$$Z = \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \tag{4}$$

Model Selection and testing hypotheses. Let us consider now M possible models (or hypotheses), $\mathcal{M}_1, \dots, \mathcal{M}_M$, with prior probability mass $p_m = \mathbb{P}(\mathcal{M}_m)$, $m = 1, \dots, M$. Note that, we can have variables of interest $\boldsymbol{\theta}^{(m)} = [\theta_1^{(m)}, \theta_2^{(m)}, \dots, \theta_{D_m}^{(m)}] \in \Theta_m \in \mathbb{R}^{D_m}$, with possibly different dimensions in the different models. The posterior of the m -th model is given by

$$p(\mathcal{M}_m|\mathbf{y}) = \frac{p_m p(\mathbf{y}|\mathcal{M}_m)}{p(\mathbf{y})} \propto p_m Z_m \tag{5}$$

where $Z_m = p(\mathbf{y}|\mathcal{M}_m) = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta}_m, \mathcal{M}_m)g(\boldsymbol{\theta}_m|\mathcal{M}_m)d\boldsymbol{\theta}_m$, and $p(\mathbf{y}) = \sum_{m=1}^M p(\mathcal{M}_m)p(\mathbf{y}|\mathcal{M}_m)$. Moreover, the ratio of two marginal likelihoods

$$\frac{Z_m}{Z_{m'}} = \frac{p(\mathbf{y}|\mathcal{M}_m)}{p(\mathbf{y}|\mathcal{M}_{m'})} = \frac{p(\mathcal{M}_m|\mathbf{y})/p_m}{p(\mathcal{M}_{m'}|\mathbf{y})/p_{m'}}, \tag{6}$$

also known as *Bayes factors*, represents the posterior to prior odds of models m and m' . If some quantity of interest is common to all models, the posterior of this quantity can be studied via *model averaging* [19], i.e., a complete posterior distribution as a mixture of M partial posteriors linearly combined with weights proportionally to $p(\mathcal{M}_m|\mathbf{y})$ (see, e.g, [20, 21]). Therefore, in all these scenarios, we need the computation of Z_m for all $m = 1, \dots, M$. In this work, we describe different computational techniques for calculating Z_m , mostly based on Markov Chain Monte Carlo (MCMC) and Importance Sampling (IS) algorithms [2]. Hereafter, we assume proper prior $g(\boldsymbol{\theta}|\mathcal{M}_m)$. Regarding the use of *improper priors* see Section 7.2. Moreover, we usually denote Z , Θ , \mathcal{M} , omitting the subindex m , to simplify notation. It is important also to remark that, in some cases, it is also necessary to approximate normalizing constants (that are also functions of the parameters) in each iteration of an MCMC algorithm, in order to allow the study of the posterior density. For instance, this is the case of the so-called double intractable posteriors [4].

Remark 1. *The evidence Z is the normalizing constant of $\pi(\boldsymbol{\theta}|\mathbf{y})$, hence most of the methods in this review can be used to approximate normalizing constants of generic pdfs.*

Remark 2. *Instead of approximating the single values Z_m for all m , another approach consists in estimating directly the ratio of two marginal likelihoods $\frac{Z_m}{Z_{m'}}$, i.e., approximating directly the Bayes factors. For these reasons, several computational methods focus on estimating the ratio of two normalizing constants. However, they can be used also for estimating a single Z_m provided that $Z_{m'}$ is known.*

Table 1: Main notation of the work.

D_θ	dimension of the parameter space, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{D_\theta}$.
D_y	Total number of data.
$\boldsymbol{\theta}$	parameters; $\boldsymbol{\theta} = [\theta_1, \dots, \theta_{D_\theta}]$.
\mathbf{y}	Data, $\mathbf{y} = [y_1, \dots, y_{D_y}]$.
$\ell(\mathbf{y} \boldsymbol{\theta})$	Likelihood function.
$g(\boldsymbol{\theta})$	Prior pdf.
$P(\boldsymbol{\theta} \mathbf{y})$	Posterior pdf, $P(\boldsymbol{\theta} \mathbf{y}) = \frac{\ell(\mathbf{y} \boldsymbol{\theta})g(\boldsymbol{\theta})}{Z}$.
$\pi(\boldsymbol{\theta} \mathbf{y})$	Unnormalized posterior, $\pi(\boldsymbol{\theta} \mathbf{y}) = \ell(\mathbf{y} \boldsymbol{\theta})g(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta} \mathbf{y})$.
$Z = p(\mathbf{y})$	Marginal likelihood, a.k.a., Bayesian evidence $Z = \int_\Theta \pi(\boldsymbol{\theta} \mathbf{y})d\boldsymbol{\theta}$.
$\bar{q}(\boldsymbol{\theta})$	Proposal pdf.
$q(\boldsymbol{\theta})$	Unnormalized proposal function, $q(\boldsymbol{\theta}) \propto \bar{q}(\boldsymbol{\theta})$.

2.2 Model fit and model complexity

2.2.1 Bounds of the evidence Z

Let us denote the maximum and minimum value of the likelihood function as $\ell_{\min} = \ell(\mathbf{y}|\boldsymbol{\theta}_{\min}) = \min_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{y}|\boldsymbol{\theta})$, and $\ell_{\max} = \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}) = \max_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{y}|\boldsymbol{\theta})$, respectively. Note that

$$Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} \leq \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}) \int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}).$$

Similarly, we can obtain $Z \geq \ell(\mathbf{y}|\boldsymbol{\theta}_{\min})$. The maximum and minimum value of Z are reached with two degenerate choices of the prior, $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\max})$ and $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\min})$. Hence, for every other choice of $g(\boldsymbol{\theta})$, we have

$$\ell(\mathbf{y}|\boldsymbol{\theta}_{\min}) \leq Z \leq \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}). \quad (7)$$

Namely, depending on the choice of the prior $g(\boldsymbol{\theta})$, we can have any value of Bayesian evidence contained in the interval $[\ell(\mathbf{y}|\boldsymbol{\theta}_{\min}), \ell(\mathbf{y}|\boldsymbol{\theta}_{\max})]$. For further discussion see Section 7.

The two possible extreme values correspond to the worst and the best model fit, respectively. Below, we will see that if $Z = \ell(\mathbf{y}|\boldsymbol{\theta}_{\min})$ the chosen prior, $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\min})$, applies the greatest possible penalty to the model whereas, if $Z = \ell(\mathbf{y}|\boldsymbol{\theta}_{\max})$, the chosen prior, $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\max})$, does not apply any penalization to the model complexity (we have the maximum overfitting). Namely, the evidence Z is an average of the likelihood values, weighted according to the prior.

2.2.2 Occam factor and implicit/intrinsic complexity penalization in Z

The marginal likelihood can be expressed as

$$Z = \ell_{\max}W, \quad (8)$$

where $W \in [0, 1]$ is the *Occam factor* [22, Sect. 3]. More specifically, the Occam factor is defined as

$$W = \frac{1}{\ell_{\max}} \int_{\Theta} g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}, \quad (9)$$

and it is $\frac{\ell_{\min}}{\ell_{\max}} \leq W \leq 1$. The factor W measures the penalty of the model complexity *intrinsically* contained in the marginal likelihood Z : this penalization depends on the chosen prior and the number of data involved. We show below that the Occam factor measures the “overlap” between likelihood and prior, i.e., how diffuse the prior is with respect to the likelihood function. Finally, it is important to remark that, considering the posterior of the m -th model $p(\mathcal{M}_m|\mathbf{y})$, we have another possible penalization term due to the prior $p_m = P(\mathcal{M}_m) \in [0, 1]$, i.e.,

$$p(\mathcal{M}_m|\mathbf{y}) \propto Zp_m = \ell_{\max}Wp_m = \ell_{\max}\widetilde{W},$$

where we have defined the *posterior Occam factor* as $\widetilde{W} = Wp_m$.

2.2.3 Occam factor with uniform priors

One-dimensional case. Let start with a single parameter, $\boldsymbol{\theta} = \theta$, and a uniform prior in $[a, b]$. We can define the amount of the likelihood mass is contained inside the prior bounds,

$$\Delta_\ell = \frac{1}{\ell_{\max}} \int_a^b \ell(\mathbf{y}|\theta) d\theta, \quad \text{where} \quad \ell_{\max} = \max_{\theta \in \Theta} \ell(\mathbf{y}|\theta). \quad (10)$$

Defining also the width of the prior as $\Delta_\theta = |\Theta| = b - a$, note that $0 \leq \Delta_\ell \leq \Delta_\theta$, where the equality $\Delta_\ell = \Delta_\theta$ is given when the likelihood is $\ell(\mathbf{y}|\theta) = \ell_{\max}$ is constant. The Occam factor is given as the ratio of Δ_ℓ and the width of a uniform prior Δ_θ [22],

$$W = \frac{\Delta_\ell}{\Delta_\theta}. \quad (11)$$

If the likelihood function is integrable in \mathbb{R} , then there exists a finite upper bound for Δ_ℓ when $\Delta_\theta \rightarrow \infty$, that is $\Delta_\ell^* = \frac{1}{\ell_{\max}} \int_{-\infty}^{+\infty} \ell(\mathbf{y}|\theta) d\theta$. Hence, in this scenario, we can see that an increase of Δ_θ makes that W approaches 0.

Multidimensional case. Consider now a multidimensional case, $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_{D_\theta}] \in \Theta \subseteq \mathbb{R}^{D_\theta}$, where we can use the same uniform prior, with the same width $\Delta_\theta = |\Theta|$, for all the parameters. In this case, $\Delta_\ell = \frac{1}{\ell_{\max}} \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \leq (\Delta_\theta)^{D_\theta}$ is D_θ -dimensional integral, and $\ell_{\max} = \max \ell(\mathbf{y}|\boldsymbol{\theta})$. Then, for D_θ parameters, the Occam factor is

$$W = \frac{\Delta_\ell}{(\Delta_\theta)^{D_\theta}}. \quad (12)$$

Usually, as D_θ grows, the fitting improves until reaching (or approaching) a maximum, possible overfitting. Then, with D_θ big enough, ℓ_{\max} tends to be virtually constant (reaching the maximum overfitting). If $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$ grows slower than $(\Delta_\theta)^{D_\theta}$ as $D_\theta \rightarrow \infty$, and assuming for an illustrative purpose $\Delta_\theta > 1$, then W converges to 0 as $D_\theta \rightarrow \infty$. That is, when we introduce more and more parameters, the increase in model fit will be dominated, at some point, by the model complexity penalization implicitly contained in the evidence Z .

2.2.4 Marginal likelihood and information criteria

Considering the expressions (8) and (12) and taking the logarithm, we obtain

$$\begin{aligned} \log Z &= \log \ell_{\max} + \log W = \log \ell_{\max} + \log \Delta_\ell - D_\theta \log \Delta_\theta, \\ &= \log \ell_{\max} + \eta D_\theta, \end{aligned} \quad (13)$$

where $\eta = \frac{\log \Delta_\ell}{D_\theta} - \log \Delta_\theta$ is a constant value, which also depends on the number of data D_y and, generally, $\eta = \eta(D_y, D_\theta)$. Different model selection rules in the literature consider the simplification $\eta = \eta(D_y)$. Note that $\log \ell_{\max}$ is a fitting term whereas ηD_θ is a penalty for the model complexity. Instead of maximizing Z (or $\log Z$) for model selection purposes, several authors consider the minimization of some cost functions derived by different information

criteria. To connect them with the marginal likelihood maximization, we consider the expression of $-2 \log Z = -2I$ where $I = -\log Z$ resembles the Shannon information associated to $Z = p(\mathbf{y})$, i.e.,

$$2I = -2 \log Z = -2 \log \ell_{\max} - 2\eta D_{\theta}. \quad (14)$$

The expression above encompasses several well-known information criteria proposed in the literature and shown in Table 2, which differ for the choice of η . In all these cases, η is just a function of the number of data D_y . More details regarding these information criteria are given in Section 3.

Remark 3. *The penalty term in the information criteria is the same for every parameter. The Bayesian approach allows the choice of different penalties, assuming different priors, one for each parameter.*

Table 2: Different information criterion for model selection.

Criterion	Choice - approximation of η
Bayesian-Schwarz information criterion (BIC) [23]	$-\frac{1}{2} \log D_y$
Akaike information criterion (AIC) [9]	-1
Hannan-Quinn information criterion (HQIC) [24]	$-\log(\log(D_y))$

2.3 A general overview of the computational methods

After a depth revision of the literature, we have recognized four main families of techniques, described below. We list them in order of complexity, from the simplest to the most complex underlying main idea. However, each class can contain both simple and very sophisticated algorithms.

Family 1: Deterministic approximations. These methods consider an analytical approximation of the function $P(\boldsymbol{\theta}|\mathbf{y})$. The Laplace method and the Bayesian Information Criterion (BIC), belongs to this family (see Section 3).

Family 2: Methods based on density estimation. This class of algorithms uses the equality

$$\hat{Z} = \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})}{\hat{P}(\boldsymbol{\theta}^*|\mathbf{y})}, \quad (15)$$

where $\hat{P}(\boldsymbol{\theta}^*|\mathbf{y}) \approx P(\boldsymbol{\theta}^*|\mathbf{y})$ represents an estimation of the density $P(\boldsymbol{\theta}|\mathbf{y})$ at some point $\boldsymbol{\theta}^*$. Generally, the point $\boldsymbol{\theta}^*$ is chosen in a high-probability region. The techniques in this family differ in the procedure employed for obtaining the estimation $\hat{P}(\boldsymbol{\theta}^*|\mathbf{y})$. One famous example is the Chib's method [25]. Section 3 is devoted to describe methods belonging to family 1 and family 2.

Family 3: *Importance sampling (IS) schemes.* The IS methods are based on rewriting Eq. (2) as an expected value w.r.t. a simpler normalized density $\bar{q}(\boldsymbol{\theta})$, i.e., $Z = \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = E_{\bar{q}} \left[\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\bar{q}(\boldsymbol{\theta})} \right]$. This is the most considered class of methods in the literature, containing numerous variants, extensions and generalizations. We devote Sections 4-5 to this family of techniques.

Family 4: *Methods based on a vertical representation.* These schemes rely on changing the expression of $Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$ (that is a multidimensional integral) to equivalent one-dimensional integrals [26, 27, 28]. Then, a quadrature scheme is applied to approximate this one-dimensional integral. The most famous example is the nested sampling algorithm [28]. Section 6 is devoted to this class of methods.

3 Methods based on deterministic approximations and density estimation

In this section, we consider approximations of $P(\boldsymbol{\theta}|\mathbf{y})$, or its unnormalized version $\pi(\boldsymbol{\theta}|\mathbf{y})$, in order to obtain an estimation Z . In a first approach, the methods consider $P(\boldsymbol{\theta}|\mathbf{y})$ or $\pi(\boldsymbol{\theta}|\mathbf{y})$ as a function, and try to obtain a good approximation given another parametric or non-parametric family of functions. Another approach consists in approximating $P(\boldsymbol{\theta}|\mathbf{y})$ only at one specific point $\boldsymbol{\theta}^*$, i.e., $\hat{P}(\boldsymbol{\theta}^*|\mathbf{y}) \approx P(\boldsymbol{\theta}^*|\mathbf{y})$ ($\boldsymbol{\theta}^*$ is usually chosen in high posterior probability regions), and then using the identity

$$\hat{Z} = \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})}{\hat{P}(\boldsymbol{\theta}^*|\mathbf{y})}. \quad (16)$$

The latter scheme is often called *candidate's estimation*.

3.1 Laplace's method

Let us define $\hat{\boldsymbol{\theta}}_{\text{MAP}} \approx \boldsymbol{\theta}_{\text{MAP}} = \arg \max P(\boldsymbol{\theta}|\mathbf{y})$ (obtained by some optimization method), which is an approximation of the *maximum a posteriori* (MAP), and consider a Gaussian approximation of $P(\boldsymbol{\theta}|\mathbf{y})$ around $\hat{\boldsymbol{\theta}}_{\text{MAP}}$, i.e.,

$$\hat{P}(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{\text{MAP}}, \hat{\boldsymbol{\Sigma}}), \quad (17)$$

with $\hat{\boldsymbol{\Sigma}} \approx -\mathbf{H}^{-1}$, which is an approximation of the negative inverse Hessian matrix of $\log \pi(\boldsymbol{\theta}|\mathbf{y})$ at $\hat{\boldsymbol{\theta}}_{\text{MAP}}$. Replacing in Eq. (16), with $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_{\text{MAP}}$, we obtain the Laplace approximation

$$\hat{Z} = \frac{\pi(\hat{\boldsymbol{\theta}}_{\text{MAP}}|\mathbf{y})}{\mathcal{N}(\hat{\boldsymbol{\theta}}_{\text{MAP}}|\hat{\boldsymbol{\theta}}_{\text{MAP}}, \hat{\boldsymbol{\Sigma}})} = (2\pi)^{\frac{D_x}{2}} |\hat{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \pi(\hat{\boldsymbol{\theta}}_{\text{MAP}}|\mathbf{y}). \quad (18)$$

This is equivalent to the classical derivation of Laplace's estimator, which is based on expanding the $\log \pi(\boldsymbol{\theta}|\mathbf{y}) = \log(\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}))$ as quadratic around $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ and substituting in $Z = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$,

that is,

$$Z = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \int \exp\{\log \pi(\boldsymbol{\theta}|\mathbf{y})\}d\boldsymbol{\theta} \quad (19)$$

$$\approx \int \exp \left\{ \log \pi(\hat{\boldsymbol{\theta}}_{\text{MAP}}|\mathbf{y}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{MAP}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{MAP}}) \right\} d\boldsymbol{\theta} \quad (20)$$

$$= (2\pi)^{\frac{D_\theta}{2}} |\hat{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \pi(\hat{\boldsymbol{\theta}}_{\text{MAP}}|\mathbf{y}). \quad (21)$$

In [29], they propose to use samples generated by a Metropolis-Hastings algorithm to estimate the quantities $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ and $\hat{\boldsymbol{\Sigma}}$ [2]. The resulting method is called Laplace-Metropolis estimator. The authors in [30] present different variants of the Laplace's estimator. A relevant extension for Gaussian Markov random field models, is the so-called *integrated nested Laplace approximation* (INLA) [31].

3.2 Bayesian-Schwarz information criterion (BIC)

Let us define $\hat{\boldsymbol{\theta}}_{\text{MLE}} \approx \boldsymbol{\theta}_{\text{MLE}} = \arg \max \ell(\mathbf{y}|\boldsymbol{\theta})$. The following quantity

$$\text{BIC} = D_\theta \log D_y - 2 \log \ell(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{MLE}}), \quad (22)$$

was introduced by Gideon E. Schwarz in [23], where D_θ represents the number of parameters of the model ($\boldsymbol{\theta} \in \mathbb{R}^{D_\theta}$), D_y is the number of data,¹ and $\ell(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{MLE}})$ is the estimated maximum value of the likelihood function. The value of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ can be obtained using samples generated by a MCMC scheme. The BIC expression can be derived similarly to the Laplace's method, but this time with a second-order Taylor expansion of the log Z around its maximum $\boldsymbol{\theta}_{\text{MLE}}$ and a first-order expansion of the prior around $\boldsymbol{\theta}_{\text{MLE}}$ [32, Ch. 9.1.3]. The derivation is given in the Supplementary Material. Then, the final approximation is

$$Z \approx \hat{Z} = \exp \left(\log \ell(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{MLE}}) - \frac{D_\theta}{2} \log D_y \right) = \exp \left(-\frac{1}{2} \text{BIC} \right), \quad \text{as } D_y \rightarrow \infty, \quad (23)$$

and $\text{BIC} \approx -2 \log Z$, asymptotically as the number of data D_y grows. Then, smaller BIC values are associated to better models. Note that BIC clearly takes into account the complexity of the model since higher BIC values are given to models with more number of parameters D_θ . Namely the penalty $D_\theta \log D_y$ discourages overfitting, since increasing the number of parameters generally improves the goodness of the fit. Other criteria can be found in the literature, such as the well-known Akaike information criterion (AIC),

$$\text{AIC} = 2D_\theta - 2 \log \ell(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{MLE}}).$$

However, they are not an approximation of the marginal likelihood Z and are usually founded on information theory derivations. Generally, they have the form of $c_p - 2 \log \ell(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{MLE}})$ where the

¹Note that, for simplicity, we are considering scalar observations y_i , so that the dimension D_y of the data vector \mathbf{y} coincides with the number of data.

penalty term c_p of the model complexity changes in each different criterion (e.g., $c_p = D_\theta \log D_y$ in BIC and $c_p = 2D_\theta$ in AIC). Another example that uses MCMC samples is the Deviance Information Criterion (DIC), i.e.,

$$\text{DIC} = -\frac{4}{N} \sum_{n=1}^N \log \ell(\mathbf{y}|\boldsymbol{\theta}_n) - 2 \log \ell(\mathbf{y}|\bar{\boldsymbol{\theta}}), \quad \text{where} \quad \bar{\boldsymbol{\theta}} = \frac{1}{N} \sum_{n=1}^N \boldsymbol{\theta}_n, \quad (24)$$

and $\{\boldsymbol{\theta}_n\}_{n=1}^N$ are outputs of an MCMC algorithm [9]. In this case, note that $c_p = -\frac{4}{N} \sum_{i=1}^N \log \ell(\mathbf{y}|\boldsymbol{\theta}_i)$. DIC is considered more adequate for hierarchical models than AIC, BIC [9], but is not directly related to the marginal likelihood [11]. See also related comments in Section 2.2.4.

3.3 Kernel density estimation (KDE)

KDE can be used to approximate the value of the posterior density at a given point $\boldsymbol{\theta}^*$, and then consider $Z \approx \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})}{\hat{P}(\boldsymbol{\theta}^*|\mathbf{y})}$. For instance, we can build a kernel density estimate (KDE) of $P(\boldsymbol{\theta}|\mathbf{y})$ based on M samples distributed according to the posterior (obtained via an MCMC algorithm, for instance) by using M normalized kernel functions $k(\boldsymbol{\theta}|\boldsymbol{\mu}_m, h)$ (with $\int_{\Theta} k(\boldsymbol{\theta}|\boldsymbol{\mu}_m, h) d\boldsymbol{\theta} = 1$ for all m) where $\boldsymbol{\mu}_m$ is a location parameter and h is a scale parameter,

$$\hat{P}(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{1}{M} \sum_{m=1}^M k(\boldsymbol{\theta}^*|\boldsymbol{\mu}_m, h), \quad \{\boldsymbol{\mu}_m\}_{m=1}^M \sim P(\boldsymbol{\theta}|\mathbf{y}) \quad (\text{e.g., via MCMC}). \quad (25)$$

Generally, $\hat{P}(\boldsymbol{\theta}^*|\mathbf{y})$ is a biased estimation of $P(\boldsymbol{\theta}^*|\mathbf{y})$. The estimator is $\hat{Z} = \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})}{\hat{P}(\boldsymbol{\theta}^*|\mathbf{y})}$ where the point $\boldsymbol{\theta}^*$ can be chosen as $\hat{\boldsymbol{\theta}}_{\text{MAP}}$. If we consider N different points $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ (selected without any specific rule) we can also write a more general approximation,

$$\hat{Z} = \frac{1}{N} \sum_{n=1}^N \frac{\pi(\boldsymbol{\theta}_n|\mathbf{y})}{\hat{P}(\boldsymbol{\theta}_n|\mathbf{y})}. \quad (26)$$

Remark 4. *The estimator above is generally biased and depends on the choices of (a) of the points $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$, (b) the scale parameter h , and (c) the number of samples M for building $\hat{P}(\boldsymbol{\theta}^*|\mathbf{y})$.*

Remark 5. *A improved version of this approximation can be obtained by the importance sampling approach described in Sect. 4, where $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ are drawn from the KDE mixture $\hat{P}(\boldsymbol{\theta}|\mathbf{y})$. In this case, the resulting estimator is unbiased.*

3.4 Chib's method

In [25, 33], the authors present more sophisticated methods to estimate $P(\boldsymbol{\theta}^*|\mathbf{y})$ using outputs from Gibbs sampling and the Metropolis-Hastings (MH) algorithm respectively [2]. Here we only present the latter method, since it can be applied in more general settings. In [33], the authors

propose to estimate the value of the posterior at one point $\boldsymbol{\theta}^*$, i.e., $P(\boldsymbol{\theta}^*|\mathbf{y})$, using the output from a MH sampler. More specifically, let us denote the current state as $\boldsymbol{\theta}$. A possible candidate as future state $\mathbf{z} \sim \varphi(\mathbf{z}|\boldsymbol{\theta})$ (where $\varphi(\mathbf{z}|\boldsymbol{\theta})$ represents the proposal density used within MH), is accepted with probability $\alpha(\boldsymbol{\theta}, \mathbf{z}) = \min \left\{ 1, \frac{\pi(\mathbf{z}|\mathbf{y})\varphi(\boldsymbol{\theta}|\mathbf{z})}{\pi(\boldsymbol{\theta}|\mathbf{y})\varphi(\mathbf{z}|\boldsymbol{\theta})} \right\}$ [2, 34]. This is just an example of $\alpha(\boldsymbol{\theta}, \mathbf{z})$ that by construction the probability α satisfies the detailed balance condition [34, Section 2.4],[35], i.e.,

$$\alpha(\boldsymbol{\theta}, \mathbf{z})\varphi(\mathbf{z}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{y}) = \alpha(\mathbf{z}, \boldsymbol{\theta})\varphi(\boldsymbol{\theta}|\mathbf{z})P(\mathbf{z}|\mathbf{y}). \quad (27)$$

By integrating in $\boldsymbol{\theta}$ both sides, we obtain

$$\begin{aligned} \int_{\Theta} \alpha(\boldsymbol{\theta}, \mathbf{z})\varphi(\mathbf{z}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} &= \int_{\Theta} \alpha(\mathbf{z}, \boldsymbol{\theta})\varphi(\boldsymbol{\theta}|\mathbf{z})P(\mathbf{z}|\mathbf{y})d\boldsymbol{\theta}, \\ &= P(\mathbf{z}|\mathbf{y}) \int_{\Theta} \alpha(\mathbf{z}, \boldsymbol{\theta})\varphi(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta}, \end{aligned}$$

hence finally we can solve with respect to $P(\mathbf{z}|\mathbf{y})$ obtaining

$$P(\mathbf{z}|\mathbf{y}) = \frac{\int_{\Theta} \alpha(\boldsymbol{\theta}, \mathbf{z})\varphi(\mathbf{z}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}}{\int_{\Theta} \alpha(\mathbf{z}, \boldsymbol{\theta})\varphi(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta}}. \quad (28)$$

This suggests the following estimate of $P(\boldsymbol{\theta}^*|\mathbf{y})$ at a specific point $\boldsymbol{\theta}^*$ (note that $\boldsymbol{\theta}^*$ plays the role of \mathbf{z} in the equation above),

$$\widehat{P}(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{\frac{1}{N_1} \sum_{i=1}^{N_1} \alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*)\varphi(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)}{\frac{1}{N_2} \sum_{j=1}^{N_2} \alpha(\boldsymbol{\theta}^*, \mathbf{v}_j)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim P(\boldsymbol{\theta}|\mathbf{y}), \quad \{\mathbf{v}_j\}_{j=1}^{N_2} \sim \varphi(\boldsymbol{\theta}|\boldsymbol{\theta}^*). \quad (29)$$

The same outputs of the MH scheme can be considered as $\{\boldsymbol{\theta}_i\}_{i=1}^{N_1}$. The final estimator is again $\widehat{Z} = \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})}{\widehat{P}(\boldsymbol{\theta}^*|\mathbf{y})}$, i.e.,

$$\widehat{Z} = \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y}) \frac{1}{N_2} \sum_{j=1}^{N_2} \alpha(\boldsymbol{\theta}^*, \mathbf{v}_j)}{\frac{1}{N_1} \sum_{i=1}^{N_1} \alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*)\varphi(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim P(\boldsymbol{\theta}|\mathbf{y}), \quad \{\mathbf{v}_j\}_{j=1}^{N_2} \sim \varphi(\boldsymbol{\theta}|\boldsymbol{\theta}^*). \quad (30)$$

The point $\boldsymbol{\theta}^*$ is usually chosen in an high probability region. Interesting discussions are contained in [36], where the authors also show that this estimator is related to bridge sampling idea described in Section 4.2. For more details, see Section 4.2.2.

3.5 Interpolative approaches

Another possibility is to approximate Z by substituting the true $\pi(\boldsymbol{\theta}|\mathbf{y})$ with interpolation or a regression function $\widehat{\pi}(\boldsymbol{\theta}|\mathbf{y})$ in the integral (4). For simplicity, we focus on the interpolation case, but all the considerations can be easily extended for a regression scenario. Given a set of nodes

$\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N\} \subset \Theta$ and N nonlinear functions $k(\boldsymbol{\theta}, \boldsymbol{\theta}') : \Theta \times \Theta \rightarrow \mathbb{R}$ chosen in advance by the user (generally, centered around $\boldsymbol{\theta}'$), we can build the interpolant of unnormalized posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ as follows

$$\widehat{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^N \beta_i k(\boldsymbol{\theta}, \boldsymbol{\theta}_i), \quad (31)$$

where $\beta_i \in \mathbb{R}$ and the subindex u denotes that is an approximation of the unnormalized function $\pi(\boldsymbol{\theta}|\mathbf{y})$. The coefficients β_i are chosen such that $\widehat{\pi}_u(\boldsymbol{\theta}|\mathbf{y})$ interpolates the points $\{\boldsymbol{\theta}_n, \pi(\boldsymbol{\theta}_n|\mathbf{y})\}$, that is, $\widehat{\pi}(\boldsymbol{\theta}_n|\mathbf{y}) = \pi(\boldsymbol{\theta}_n|\mathbf{y})$. Then, we desire that

$$\sum_{i=1}^N \beta_i k(\boldsymbol{\theta}_n, \boldsymbol{\theta}_i) = \pi(\boldsymbol{\theta}_n|\mathbf{y}),$$

for all $n = 1, \dots, N$. Hence, we can write a $N \times N$ linear system where the β_i are the N unknowns, i.e.,

$$\begin{pmatrix} k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) & k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & \dots & k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_N) \\ k(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1) & k(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2) & \dots & k(\boldsymbol{\theta}_2, \boldsymbol{\theta}_N) \\ \vdots & & \ddots & \vdots \\ k(\boldsymbol{\theta}_N, \boldsymbol{\theta}_1) & k(\boldsymbol{\theta}_N, \boldsymbol{\theta}_2) & \dots & k(\boldsymbol{\theta}_N, \boldsymbol{\theta}_N) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix} = \begin{pmatrix} \pi(\boldsymbol{\theta}_1|\mathbf{y}) \\ \pi(\boldsymbol{\theta}_2|\mathbf{y}) \\ \vdots \\ \pi(\boldsymbol{\theta}_N|\mathbf{y}) \end{pmatrix} \quad (32)$$

In matrix form, we have

$$\mathbf{K}\boldsymbol{\beta} = \mathbf{y}, \quad (33)$$

where $(\mathbf{K})_{i,j} = k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and $\mathbf{y} = [\pi(\boldsymbol{\theta}_1|\mathbf{y}), \dots, \pi(\boldsymbol{\theta}_N|\mathbf{y})]^\top$. Thus, the solution is $\boldsymbol{\beta} = \mathbf{K}^{-1}\mathbf{y}$. Now the interpolant $\widehat{\pi}_u(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^N \beta_i k(\boldsymbol{\theta}, \boldsymbol{\theta}_i)$ can be used to approximate Z as follows

$$\widehat{Z} = \int_{\Theta} \widehat{\pi}_u(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \sum_{i=1}^N \beta_i \int_{\Theta} k(\boldsymbol{\theta}, \boldsymbol{\theta}_i) d\boldsymbol{\theta}. \quad (34)$$

If we are able to compute analytically $\int_{\Theta} k(\boldsymbol{\theta}, \boldsymbol{\theta}_i) d\boldsymbol{\theta}$, we have an approximation \widehat{Z} . Some suitable choices of $k(\cdot, \cdot)$ are rectangular, triangular and Gaussian functions. More specifically, if all the nonlinearities $k(\boldsymbol{\theta}, \boldsymbol{\theta}_i)$ are normalized (i.e. $\int_{\Theta} k(\boldsymbol{\theta}, \boldsymbol{\theta}_i) d\boldsymbol{\theta} = 1$), the approximation of Z is $\widehat{Z} = \sum_{i=1}^N \beta_i$. This approach is related to the so-called Bayesian quadrature (using Gaussian process approximation) [37] and the sticky proposal constructions within MCMC or rejection sampling algorithms [38, 39, 40, 41]. Adaptive schemes adding sequentially more nodes could be also considered, improving the approximation \widehat{Z} [39, 40]. The quality of the interpolating approximation deteriorates as the dimension of $\boldsymbol{\theta}$ grows (see e.g. [42] for explicit error bounds).

4 Techniques based on IS

Most of the techniques for approximating the marginal likelihood are based on the importance sampling (IS) approach. Other methods are directly or indirectly related to the IS framework. In

this sense, this section is the core of this survey. The standard IS scheme relies on the following equality,

$$Z = \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \mathbb{E}_{\bar{q}} \left[\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\bar{q}(\boldsymbol{\theta})} \right] = \int_{\Theta} \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\bar{q}(\boldsymbol{\theta})} \bar{q}(\boldsymbol{\theta})d\boldsymbol{\theta} \quad (35)$$

$$= \int_{\Theta} \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{\bar{q}(\boldsymbol{\theta})} \bar{q}(\boldsymbol{\theta})d\boldsymbol{\theta}, \quad (36)$$

where $\bar{q}(\boldsymbol{\theta})$ is a simpler normalized proposal density, $\int_{\Theta} \bar{q}(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$.

IS version 1. Drawing N independent samples from proposal $\bar{q}(\boldsymbol{\theta})$, the *unbiased* IS estimator (denoted as IS vers-1) of Z is

$$\widehat{Z}_{IS1} = \frac{1}{N} \sum_{i=1}^N \frac{\pi(\boldsymbol{\theta}_i|\mathbf{y})}{\bar{q}(\boldsymbol{\theta}_i)} \quad (37)$$

$$= \frac{1}{N} \sum_{i=1}^N w_i, \quad (38)$$

$$= \frac{1}{N} \sum_{i=1}^N \frac{\ell(\mathbf{y}|\boldsymbol{\theta}_i)g(\boldsymbol{\theta}_i)}{\bar{q}(\boldsymbol{\theta}_i)} = \frac{1}{N} \sum_{i=1}^N \rho_i \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}(\boldsymbol{\theta}), \quad (39)$$

where $w_i = \frac{\pi(\boldsymbol{\theta}_i|\mathbf{y})}{\bar{q}(\boldsymbol{\theta}_i)}$ are the standard IS weights and $\rho_i = \frac{g(\boldsymbol{\theta}_i)}{\bar{q}(\boldsymbol{\theta}_i)}$.

Optimal proposal in IS vers-1. The optimal proposal, in terms of mean square error (MSE), in the standard IS scheme above is $\bar{q}^{\text{opt}}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$.

IS version 2. An alternative IS estimator (denoted as IS vers-2) is given by, considering a possibly unnormalized proposal pdf $q(\boldsymbol{\theta}) \propto \bar{q}(\boldsymbol{\theta})$ (the case $q(\boldsymbol{\theta}) = \bar{q}(\boldsymbol{\theta})$ is also included),

$$\widehat{Z}_{IS2} = \frac{1}{\sum_{n=1}^N \frac{g(\boldsymbol{\theta}_n)}{q(\boldsymbol{\theta}_n)}} \sum_{i=1}^N \frac{g(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)} \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad (40)$$

$$= \frac{1}{\sum_{n=1}^N \rho_n} \sum_{i=1}^N \rho_i \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad (41)$$

$$= \sum_{i=1}^N \bar{\rho}_i \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}(\boldsymbol{\theta}). \quad (42)$$

The estimator above is biased. However, it is a convex combination of likelihood values $\ell(\mathbf{y}|\boldsymbol{\theta}_i)$ since $\sum_{i=1}^N \bar{\rho}_i = 1$. Hence, in this case $\min_i \ell(\mathbf{y}|\boldsymbol{\theta}_i) \leq \widehat{Z} \leq \max_i \ell(\mathbf{y}|\boldsymbol{\theta}_i)$, i.e., the estimator fulfills the bounds of Z , shown Section 2.2. Moreover, the estimator allows the use of an unnormalized proposal pdf $q(\boldsymbol{\theta}) \propto \bar{q}(\boldsymbol{\theta})$ and $\rho_i = \frac{g(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}$. For instance, one could consider $\bar{q}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$, i.e.,

generate samples $\{\boldsymbol{\theta}_i\}_{i=1}^N \sim P(\boldsymbol{\theta}|\mathbf{y})$ by an MCMC algorithm and then evaluate $\rho_i = \frac{g(\boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i|\mathbf{y})}$.

Optimal proposal in IS vers-2. The optimal proposal, in terms of MSE, for the IS vers-2 is $\bar{q}^{\text{opt}}(\boldsymbol{\theta}) \propto |P(\boldsymbol{\theta}|\mathbf{y}) - g(\boldsymbol{\theta})|$.

Table 3 summarizes the IS estimators and shows some important special cases that will be described in the next section.

Table 3: IS estimators Eqs. (37)-(40) and relevant special cases.

$\widehat{Z}_{IS1} = \frac{1}{N} \sum_{i=1}^N \frac{g(\boldsymbol{\theta}_i)}{\bar{q}(\boldsymbol{\theta}_i)} \ell(\mathbf{y} \boldsymbol{\theta}_i) = \frac{1}{N} \sum_{i=1}^N \rho_i \ell(\mathbf{y} \boldsymbol{\theta}_i), \quad \rho_i = \frac{g(\boldsymbol{\theta}_i)}{\bar{q}(\boldsymbol{\theta}_i)}$					
Name	Estimator	$q(\boldsymbol{\theta})$	$\bar{q}(\boldsymbol{\theta})$	Need of MCMC	Unbiased
Naive Monte Carlo	$\frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y} \boldsymbol{\theta}_i)$	$g(\boldsymbol{\theta})$	$g(\boldsymbol{\theta})$	—	✓
$\widehat{Z}_{IS2} = \frac{1}{\sum_{n=1}^N \frac{g(\boldsymbol{\theta}_n)}{\bar{q}(\boldsymbol{\theta}_n)}} \sum_{i=1}^N \frac{g(\boldsymbol{\theta}_i)}{\bar{q}(\boldsymbol{\theta}_i)} \ell(\mathbf{y} \boldsymbol{\theta}_i) = \sum_{i=1}^N \bar{\rho}_i \ell(\mathbf{y} \boldsymbol{\theta}_i)$					
Name	Estimator	$q(\boldsymbol{\theta})$	$\bar{q}(\boldsymbol{\theta})$	Need of MCMC	Unbiased
Naive Monte Carlo	$\frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y} \boldsymbol{\theta}_i)$	$g(\boldsymbol{\theta})$	$g(\boldsymbol{\theta})$	—	✓
Harmonic mean	$\left(\frac{1}{N} \sum_{i=1}^N \frac{1}{\ell(\mathbf{y} \boldsymbol{\theta}_i)} \right)^{-1}$	$\pi(\boldsymbol{\theta} \mathbf{y})$	$P(\boldsymbol{\theta} \mathbf{y})$	✓	—

Different sub-families of IS schemes are commonly used for computing normalizing constants [43, chapter 5]. A first approach uses draws from a proposal density $\bar{q}(\boldsymbol{\theta})$ that is completely known (i.e. direct sampling and evaluate). Sophisticated choices of $\bar{q}(\boldsymbol{\theta})$ frequently imply the use of MCMC algorithms to sample from $\bar{q}(\boldsymbol{\theta})$ and that we can only evaluate $q(\boldsymbol{\theta}) \propto \bar{q}(\boldsymbol{\theta})$. The one-proposal approach is described in Section 4.1. A second class is formed by methods which use more than one proposal density or a mixture of them (see Sections 4.2, 4.3 and 5). Moreover, **adaptive importance sampling (AIS)** schemes are often designed, where the proposal (or the cloud of proposals) is improved during some iterations, in some way such that $\bar{q}_t(\boldsymbol{\theta})$ (where t is an iteration index) becomes closer and closer to the optima proposal $q^{\text{opt}}(\boldsymbol{\theta})$. For more details, see the reviews in [44]. Some AIS methods, obtained combining MCMC and IS approaches, are described in Section 5.

4.1 Techniques using draws from one proposal density

In this section, all the techniques are IS schemes which use a unique proposal pdf, and are based on the identity Eq. (35). The techniques differ in the choice of $\bar{q}(\boldsymbol{\theta})$. Recall that the optimal proposal choice for IS vers-1 is $\bar{q}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{Z} \pi(\boldsymbol{\theta}|\mathbf{y})$. This choice is clearly difficult for two reasons: (a) we have to draw from P and (b) we do not know Z , hence we cannot evaluate $\bar{q}(\boldsymbol{\theta})$

but only $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$ (where $q(\boldsymbol{\theta}) \propto \bar{q}(\boldsymbol{\theta})$). However, there are some methods based on this idea, as shown in the following. The techniques below are enumerated in an increasing order of complexity.

Naive Monte Carlo (arithmetic mean estimator). It is straightforward to note that the integral above can be expressed as $Z = \mathbb{E}_g[\ell(\mathbf{y}|\boldsymbol{\theta})]$, then we can draw N samples $\{\boldsymbol{\theta}_i\}_{i=1}^N$ from the prior $g(\boldsymbol{\theta})$ and compute the following estimator

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim g(\boldsymbol{\theta}). \quad (43)$$

Namely a simple average of the likelihoods of a sample from the prior. Note that \hat{Z} will be very inefficient (large variance) if the posterior is much more concentrated than the prior (i.e., small overlap between likelihood and prior pdfs). Therefore, alternatives have been proposed, see below. It is a special case of the IS estimator with the choice $\bar{q}(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ (i.e., the proposal pdf is the prior).

Harmonic mean (HM) estimators. The HM estimator can be directly derived from the following expected value,

$$\mathbb{E}_P \left[\frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta})} \right] = \int_{\Theta} \frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta})} P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \quad (44)$$

$$= \frac{1}{Z} \int_{\Theta} \frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta})} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{Z} \int_{\Theta} g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{Z}. \quad (45)$$

The main idea is again to use the posterior itself as proposal. Since direct sampling from $P(\boldsymbol{\theta}|\mathbf{y})$ is generally impossible, this task requires the use of MCMC algorithms. Thus, the HM estimator is

$$\hat{Z} = \left(\frac{1}{N} \sum_{i=1}^N \frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta}_i)} \right)^{-1} = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta}_i)}}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim P(\boldsymbol{\theta}|\mathbf{y}) \text{ (via MCMC)}. \quad (46)$$

The HM estimator converges almost surely to the correct value, but the variance of \hat{Z} is often high and possibly infinite.² The HM estimator is a special case of Reverse Importance Sampling (RIS) below.

Reverse Importance Sampling (RIS). The RIS scheme [45], also known as *reciprocal* IS, can be derived from the identity

$$\frac{1}{Z} = \mathbb{E}_P \left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\mathbf{y})} \right] = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\mathbf{y})} P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \quad (47)$$

²See the comments of Radford Neal's blog, <https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/>, where R. Neal defines the HM estimator as "the worst estimator ever".

where we consider an auxiliary normalized function $f(\boldsymbol{\theta})$, i.e., $\int_{\Theta} f(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$. Then, one could consider the estimator

$$\widehat{Z} = \left(\frac{1}{N} \sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i|\mathbf{y})} \right)^{-1} = \left(\frac{1}{N} \sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{\ell(\mathbf{y}|\boldsymbol{\theta}_i)g(\boldsymbol{\theta}_i)} \right)^{-1}, \quad \boldsymbol{\theta}_i \sim P(\boldsymbol{\theta}|\mathbf{y}) \text{ (via MCMC)}. \quad (48)$$

The estimator above is consistent but biased. Indeed, the expression $\frac{1}{N} \sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i|\mathbf{y})}$ is an unbiased estimator of $1/Z$, but \widehat{Z} in the Eq. (48) is not an unbiased estimator of Z . Note that $P(\boldsymbol{\theta}|\mathbf{y})$ plays the role of importance density from which we need to draw from. Therefore, another sampling technique must be used (such as a MCMC method) in order to generate samples from $P(\boldsymbol{\theta}|\mathbf{y})$. In this case, we do not need samples from $f(\boldsymbol{\theta})$, although its choice affects the precision of the approximation. Unlike in the standard IS approach, $f(\boldsymbol{\theta})$ must have lighter tails than $\pi(\boldsymbol{\theta}|\mathbf{y}) = \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})$. For further details, see the example in Section 8.1. Finally, note that the HM estimator is a special case of RIS when $f(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ in Eq. (48). In [46], the authors propose taking $f(\boldsymbol{\theta})$ that is uniform in a high posterior density region whereas, in [47], they consider taking $f(\boldsymbol{\theta})$ to be a piecewise constant function.

4.1.1 The pre-umbrella estimators

All the estimators that we have seen so far can be unified within a common formulation, considering the more general problem of estimating a ratio of two normalizing constants c_1/c_2 , where $c_i = \int q_i(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $\bar{q}_i(\boldsymbol{\theta}) = q_i(\boldsymbol{\theta})/c_i$, $i = 1, 2$. Assuming we can evaluate both $q_1(\boldsymbol{\theta})$, $q_2(\boldsymbol{\theta})$, and draw samples from one of them, say $\bar{q}_2(\boldsymbol{\theta})$, the importance sampling estimator of ratio c_1/c_2 is

$$\frac{c_1}{c_2} = \mathbb{E}_{\bar{q}_2} \left[\frac{q_1(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})} \right] \approx \frac{1}{N} \sum_{i=1}^N \frac{q_1(\boldsymbol{\theta}_i)}{q_2(\boldsymbol{\theta}_i)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}_2(\boldsymbol{\theta}). \quad (49)$$

Remark 6. *The relative MSE (rel-MSE) of (49), in estimation of the ratio $r = \frac{c_1}{c_2}$, i.e., $rel-MSE = \frac{\mathbb{E}[(\widehat{r}-r)^2]}{r^2}$, is given by $rel-MSE = \frac{1}{N} \chi^2(\bar{q}_1||\bar{q}_2)$, where $\chi^2(\bar{q}_1||\bar{q}_2)$ is the Pearson divergence between \bar{q}_1 and \bar{q}_2 [48].*

This framework includes almost all the estimators discussed so far in this section, as shown in Table 4. However, the IS vers-2 estimator is not a special case of Eq. (49).

Below we consider an extension of Eq. (49) where an additional density $\bar{q}_3(\boldsymbol{\theta})$ is employed for generating samples.

4.1.2 Umbrella Sampling (a.k.a. ratio importance sampling)

The IS estimator of c_1/c_2 given in Eq. (49) may be inefficient when there is little overlap between $\bar{q}_1(\boldsymbol{\theta})$ and $\bar{q}_2(\boldsymbol{\theta})$, i.e., when $\int_{\Theta} \bar{q}_1(\boldsymbol{\theta})\bar{q}_2(\boldsymbol{\theta})d\boldsymbol{\theta}$ is small. Umbrella sampling (originally proposed in the computational physics literature, [49]; also studied under the name ratio importance sampling

Table 4: Summary of techniques considering the expression (49).

Name	$q_1(\boldsymbol{\theta})$	$q_2(\boldsymbol{\theta})$	c_1	c_2	Proposal pdf $\bar{q}_2(\boldsymbol{\theta})$	c_1/c_2
IS vers-1	$\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$	Z	1	$\bar{q}(\boldsymbol{\theta})$	Z
Naive Monte Carlo	$\pi(\boldsymbol{\theta} \mathbf{y})$	$g(\boldsymbol{\theta})$	Z	1	$g(\boldsymbol{\theta})$	Z
Harmonic mean	$g(\boldsymbol{\theta})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	1	Z	$P(\boldsymbol{\theta} \mathbf{y})$	$1/Z$
RIS	$f(\boldsymbol{\theta})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	1	Z	$P(\boldsymbol{\theta} \mathbf{y})$	$1/Z$

in [48]) is based on the identity

$$\frac{c_1}{c_2} = \frac{c_1/c_3}{c_2/c_3} = \frac{\mathbb{E}_{\bar{q}_3} \left[\frac{q_1(\boldsymbol{\theta})}{q_3(\boldsymbol{\theta})} \right]}{\mathbb{E}_{\bar{q}_3} \left[\frac{q_2(\boldsymbol{\theta})}{q_3(\boldsymbol{\theta})} \right]} \approx \frac{\sum_{i=1}^N \frac{q_1(\boldsymbol{\theta}_i)}{q_3(\boldsymbol{\theta}_i)}}{\sum_{i=1}^N \frac{q_2(\boldsymbol{\theta}_i)}{q_3(\boldsymbol{\theta}_i)}}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}_3(\boldsymbol{\theta}) \quad (50)$$

where $\bar{q}_3(\boldsymbol{\theta}) \propto q_3(\boldsymbol{\theta})$ represents a *middle* density. A good choice of $\bar{q}_3(\boldsymbol{\theta})$ should have large overlaps with both $\bar{q}_i(\boldsymbol{\theta})$, $i = 1, 2$. The performance of umbrella sampling clearly depends on the choice of $\bar{q}_3(\boldsymbol{\theta})$. Note that, when $\bar{q}_3 = \bar{q}_2$, we recover Eq. (49).

Optimal umbrella proposal. The optimal umbrella sampling density $\bar{q}_3^{\text{opt}}(\boldsymbol{\theta})$, that minimizes the asymptotic relative mean-square error, is

$$\bar{q}_3^{\text{opt}}(\boldsymbol{\theta}) = \frac{|\bar{q}_1(\boldsymbol{\theta}) - \bar{q}_2(\boldsymbol{\theta})|}{\int |\bar{q}_1(\boldsymbol{\theta}') - \bar{q}_2(\boldsymbol{\theta}')| d\boldsymbol{\theta}'} = \frac{|q_1(\boldsymbol{\theta}) - \frac{c_1}{c_2} q_2(\boldsymbol{\theta})|}{\int |q_1(\boldsymbol{\theta}') - \frac{c_1}{c_2} q_2(\boldsymbol{\theta}')| d\boldsymbol{\theta}'}. \quad (51)$$

Remark 7. The *rel-MSE* in estimation of the ratio $\frac{c_1}{c_2}$ of the optimal umbrella estimator, with N great enough, is given by *rel-MSE* $\approx \frac{1}{N} L_1^2(\bar{q}_1, \bar{q}_2)$, where $L_1^2(\bar{q}_1, \bar{q}_2)$ denotes the L_1 -distance between \bar{q}_1 and \bar{q}_2 [48, Theorem 3.2]. Moreover, since $L_1^2(\bar{q}_1, \bar{q}_2) \leq \chi^2(\bar{q}_1 || \bar{q}_2)$, the optimal umbrella estimator is asymptotically more efficient than the estimator (49) [50, Sect. 3].

Two-stage umbrella sampling. Since this $\bar{q}_3^{\text{opt}}(\boldsymbol{\theta})$ depends on the unknown ratio $\frac{c_1}{c_2}$, it is not available for a direct use. The following two-stage procedure is often used in practice:

1. *Stage 1:* Draw N_1 samples from an arbitrary density $\bar{q}_3^{(1)}(\boldsymbol{\theta})$ and use them to obtain

$$\hat{r}^{(1)} = \frac{\sum_{i=1}^{N_1} \frac{q_1(\boldsymbol{\theta}_i)}{q_3^{(1)}(\boldsymbol{\theta}_i)}}{\sum_{i=1}^{N_1} \frac{q_2(\boldsymbol{\theta}_i)}{q_3^{(1)}(\boldsymbol{\theta}_i)}}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim \bar{q}_3^{(1)}(\boldsymbol{\theta}). \quad (52)$$

and define

$$\bar{q}_3^{(2)}(\boldsymbol{\theta}) \propto |q_1(\boldsymbol{\theta}) - \hat{r}^{(1)} q_2(\boldsymbol{\theta})|. \quad (53)$$

2. *Stage 2*: Draw N_2 samples from $\bar{q}_3^{(2)}(\boldsymbol{\theta})$ via MCMC and define the umbrella sampling estimator $\hat{r}^{(2)}$ of $\frac{c_1}{c_2}$ as follows

$$\hat{r}^{(2)} = \frac{\sum_{i=1}^{n_2} \frac{q_1(\boldsymbol{\theta}_i)}{q_3^{(2)}(\boldsymbol{\theta}_i)}}{\sum_{i=1}^{n_2} \frac{q_2(\boldsymbol{\theta}_i)}{q_3^{(2)}(\boldsymbol{\theta}_i)}}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{n_2} \sim \bar{q}_3^{(2)}(\boldsymbol{\theta}). \quad (54)$$

Remark 8. *The number of stages could be increased considering, at each t -th stage, the proposal $\bar{q}_3^{(t)}(\boldsymbol{\theta}) \propto |q_1(\boldsymbol{\theta}) - \hat{r}^{(t-1)} q_2(\boldsymbol{\theta})|$ and obtaining a new estimation $\hat{r}^{(t)}$. In this case, we have an umbrella scheme with adaptive proposal $\bar{q}_3^{(t)}(\boldsymbol{\theta})$.*

4.1.3 Umbrella for Z : the self-normalized Importance Sampling (Self-IS)

Here, we describe an important special case of the umbrella sampling approach. Considering the umbrella identity (50) an setting $q_1(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$, $q_2(\boldsymbol{\theta}) = \bar{q}_2(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$, $c_1 = Z$, $c_2 = 1$ and $c_3 \in \mathbb{R}$, we obtain

$$\hat{Z} = \frac{1}{\sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{q_3(\boldsymbol{\theta}_i)}} \sum_{i=1}^N \frac{\pi(\boldsymbol{\theta}_i|\mathbf{y})}{q_3(\boldsymbol{\theta}_i)}. \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}_3(\boldsymbol{\theta}). \quad (55)$$

which is called the *self-normalized IS* (Self-IS) estimator. Note that $f(\boldsymbol{\theta})$ is an auxiliary normalized pdf, but we draw samples from $\bar{q}_3(\boldsymbol{\theta})$. In order to understand the reason of its name is interesting to derive it with standard IS arguments. Let us consider that our proposal $q(\boldsymbol{\theta})$ in the standard IS scheme is not normalized, and we can evaluate it up to a normalizing constant $q(\boldsymbol{\theta}) \propto \bar{q}(\boldsymbol{\theta})$. We also denote $c = \int_{\Theta} q(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Note that this also occurs in the ideal case of using $\bar{q}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{Z} \pi(\boldsymbol{\theta}|\mathbf{y})$ where $c = Z$ and $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$. In this case, we have

$$\frac{\hat{Z}}{c} = \frac{1}{N} \sum_{i=1}^N \frac{\pi(\boldsymbol{\theta}_i|\mathbf{y})}{q(\boldsymbol{\theta}_i)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}(\boldsymbol{\theta}). \quad (56)$$

Therefore, we need an additional estimation of c . We can also use IS for this goal, considering a new normalized reference function $f(\boldsymbol{\theta})$, i.e., $\int_{\Theta} f(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$. Now,

$$\frac{1}{c} = E_{\bar{q}} \left[\frac{f(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right] = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \bar{q}(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \frac{1}{N} \sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}(\boldsymbol{\theta}). \quad (57)$$

Replacing (57) into (56), we obtain the self-normalized IS estimator in Eq. (55), i.e., $\hat{Z} = \frac{1}{\sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}} \sum_{i=1}^N \frac{\pi(\boldsymbol{\theta}_i|\mathbf{y})}{q(\boldsymbol{\theta}_i)}$ with $\{\boldsymbol{\theta}_i\}_{i=1}^N \sim \bar{q}(\boldsymbol{\theta})$.

The HM estimator is also a special case of Self-IS setting again $f(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ and $\bar{q}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$, so that $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$. Moreover, the RIS estimator is a special case of the Self-IS estimator above when $\bar{q}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$ and $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$.

Optimal self-IS (O-Self-IS). Since the Self-IS estimator is a special case of umbrella sampling, the optimal proposal in this case is $\bar{q}^{\text{opt}}(\boldsymbol{\theta}) \propto |P(\boldsymbol{\theta}|\mathbf{y}) - f(\boldsymbol{\theta})|$, and the optimal estimator is

$$\widehat{Z}_{\text{O-Self-IS}} = \frac{\sum_{i=1}^N \frac{\pi(\boldsymbol{\theta}_i)}{|P(\boldsymbol{\theta}_i|\mathbf{y}) - f(\boldsymbol{\theta}_i)|}}{\sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{|P(\boldsymbol{\theta}_i|\mathbf{y}) - f(\boldsymbol{\theta}_i)|}}, \quad \boldsymbol{\theta}_i \sim \bar{q}^{\text{opt}}(\boldsymbol{\theta}) \propto |P(\boldsymbol{\theta}|\mathbf{y}) - f(\boldsymbol{\theta})|. \quad (58)$$

Since the density cannot be evaluated (and also is not easy to draw from), this estimator is not of direct use and we need to resort to the two-stage procedure that we discussed above. Due to Remark 7, the O-Self-IS estimator is asymptotically more efficient than IS vers-1 estimator using $f(\boldsymbol{\theta})$ as proposal, i.e., drawing samples from $\bar{q}(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$.

4.1.4 Summary

The more general expressions are the two identities (49)-(50) for estimating a ratio of normalizing constants $\frac{c_1}{c_2}$. The umbrella identity (50) is the more general since three densities are involved, and contains the Eq. (49) as special case when $q_3(\boldsymbol{\theta}) = q_2(\boldsymbol{\theta})$. The Self-IS estimator coincides with the umbrella estimator when we approximate only one normalizing constant, Z (i.e., for $\frac{c_1}{c_2} = Z$). Therefore, regarding the estimation of only one constant Z , the Self-IS estimator has the more general form and includes the rest of estimators as special cases. All these connections are summarized in Table 5. Finally, Table 6 provides another summary of the one-proposal estimators of Z . Note that in the standard IS estimator the option $\bar{q}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$ is not feasible, whereas it is possible for its second version.

Table 5: Summary of techniques considering the umbrella sampling identity (50) for computing $\frac{c_1}{c_2} = Z$. Note that Self-IS has the more general form and includes the rest of estimators as special cases.

<i>For estimating a generic ratio c_1/c_2</i>							
Umbrella	$q_1(\boldsymbol{\theta})$	$q_2(\boldsymbol{\theta})$	$q_3(\boldsymbol{\theta})$	c_1	c_2	c_3	sampling from $\bar{q}_3(\boldsymbol{\theta})$
Eq. (49) - ($q_3 = q_2$)	$q_1(\boldsymbol{\theta})$	$q_2(\boldsymbol{\theta})$	$q_2(\boldsymbol{\theta})$	c_1	c_2	c_2	sampling from $\bar{q}_2(\boldsymbol{\theta})$
<i>For estimating Z</i>							
Self-IS	$\pi(\boldsymbol{\theta} \mathbf{y})$	$f(\boldsymbol{\theta})$	$q(\boldsymbol{\theta})$	Z	1	c_3	$\bar{q}(\boldsymbol{\theta})$
<i>Special cases of Self-IS</i>							
Naive Monte Carlo	$\pi(\boldsymbol{\theta} \mathbf{y})$	$g(\boldsymbol{\theta})$	$g(\boldsymbol{\theta})$			1	$g(\boldsymbol{\theta})$
Harmonic Mean	$\pi(\boldsymbol{\theta} \mathbf{y})$	$g(\boldsymbol{\theta})$	$\pi(\boldsymbol{\theta} \mathbf{y})$			Z	$P(\boldsymbol{\theta} \mathbf{y})$
RIS	$\pi(\boldsymbol{\theta} \mathbf{y})$	$f(\boldsymbol{\theta})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	Z	1	Z	$P(\boldsymbol{\theta} \mathbf{y})$
IS vers-1; Eq. (37)	$\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$	$\bar{q}(\boldsymbol{\theta})$			1	$\bar{q}(\boldsymbol{\theta})$
IS vers-2; Eq. (40)	$\pi(\boldsymbol{\theta} \mathbf{y})$	$g(\boldsymbol{\theta})$	$\bar{q}(\boldsymbol{\theta})$			1	$\bar{q}(\boldsymbol{\theta})$

In the next section, we discuss a generalization of Eq. (49) for the case where we use samples from both $\bar{q}_1(\boldsymbol{\theta})$ and $\bar{q}_2(\boldsymbol{\theta})$.

Table 6: One-proposal estimators of Z

Name	Estimator	Proposal pdf	Need of MCMC	Unbiased
IS vers-1	$\frac{1}{N} \sum_{i=1}^N \rho_i \ell(\mathbf{y} \boldsymbol{\theta}_i)$	Generic, $\bar{q}(\boldsymbol{\theta})$	—	✓
IS vers-2	$\sum_{i=1}^N \bar{\rho}_i \ell(\mathbf{y} \boldsymbol{\theta}_i)$	Generic, $\bar{q}(\boldsymbol{\theta})$	no, if $\bar{q}(\boldsymbol{\theta}) \neq P(\boldsymbol{\theta} \mathbf{y})$	—
Naive MC	$\frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y} \boldsymbol{\theta}_i)$	Prior, $g(\boldsymbol{\theta})$	—	✓
Harmonic mean	$\left(\frac{1}{N} \sum_{i=1}^N \frac{1}{\ell(\mathbf{y} \boldsymbol{\theta}_i)}\right)^{-1}$	Posterior, $P(\boldsymbol{\theta} \mathbf{y})$	✓	—
RIS	$\left(\frac{1}{N} \sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i \mathbf{y})}\right)^{-1}$	Posterior, $P(\boldsymbol{\theta} \mathbf{y})$	✓	—
Self-IS	$\left(\sum_{i=1}^N \frac{f(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}\right)^{-1} \sum_{i=1}^N \frac{\pi(\boldsymbol{\theta}_i \mathbf{y})}{q(\boldsymbol{\theta}_i)}$	Generic, $\bar{q}(\boldsymbol{\theta})$	no, if $\bar{q}(\boldsymbol{\theta}) \neq P(\boldsymbol{\theta} \mathbf{y})$	—

4.2 Techniques using draws from two proposal densities

In the previous section, we considered estimators of Z that use samples drawn from a single proposal density. More specifically, we have described several IS schemes using a generic pdf $\bar{q}(\boldsymbol{\theta})$ or $P(\boldsymbol{\theta}|\mathbf{y})$ as proposal density. In this section, we introduce schemes where $\bar{q}(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\mathbf{y})$ are employed jointly. More generally, we consider estimators of a ratio of constants, $\frac{c_2}{c_1}$, that employ samples from two proposal densities, denoted as $\bar{q}_i(\boldsymbol{\theta}) = \frac{q_i(\boldsymbol{\theta})}{c_i}$, $i = 1, 2$. Note that drawing N_1 samples from $\bar{q}_1(\boldsymbol{\theta})$ and N_2 samples from $\bar{q}_2(\boldsymbol{\theta})$ is equivalent to sampling by a *deterministic mixture* approach from the mixture $\bar{q}_{mix}(\boldsymbol{\theta}) = \frac{N_1}{N_1+N_2} \bar{q}_1(\boldsymbol{\theta}) + \frac{N_2}{N_1+N_2} \bar{q}_2(\boldsymbol{\theta})$, i.e., a single density defined as mixture of two pdfs [51]. Thus, methods drawing from a mixture of two pdfs as $\bar{q}_{mix}(\boldsymbol{\theta})$, are also considered in this section.

4.2.1 Bridge sampling identity

All the techniques, that we will describe below, are based on the following *bridge sampling* identity [52],

$$\frac{c_1}{c_2} = \frac{\mathbb{E}_{\bar{q}_2}[q_1(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})]}{\mathbb{E}_{\bar{q}_1}[q_2(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})]}. \quad (59)$$

where $\alpha(\boldsymbol{\theta})$ is an arbitrary function defined on the intersection of the supports of \bar{q}_1 and \bar{q}_2 . Note that the expression above is an extension of the Eq. (49). Indeed, taking $\alpha(\boldsymbol{\theta}) = \frac{1}{q_2(\boldsymbol{\theta})}$, we recover Eq. (49). The identity in Eq. (59) and the umbrella identity in Eq. (50) are both useful when \bar{q}_1 and \bar{q}_2 have little overlap, i.e., $\int_{\Theta} \bar{q}_1(\boldsymbol{\theta})\bar{q}_2(\boldsymbol{\theta})d\boldsymbol{\theta}$ is small. Moreover, If we set $q_1(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$, $c_1 = Z$, $q_2(\boldsymbol{\theta}) = \bar{q}(\boldsymbol{\theta})$ and $c_2 = 1$, then the identity becomes

$$Z = \frac{\mathbb{E}_{\bar{q}}[\pi(\boldsymbol{\theta}|\mathbf{y})\alpha(\boldsymbol{\theta})]}{\mathbb{E}_P[\bar{q}(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})]}. \quad (60)$$

The corresponding estimator employs samples from both \bar{q} and P , i.e.,

$$\hat{Z} = \frac{\frac{1}{N_2} \sum_{j=1}^{N_2} \alpha(\mathbf{z}_j) \pi(\mathbf{z}_j | \mathbf{y})}{\frac{1}{N_1} \sum_{i=1}^{N_1} \alpha(\boldsymbol{\theta}_i) \bar{q}(\boldsymbol{\theta}_i)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim P(\boldsymbol{\theta} | \mathbf{y}), \quad \{\mathbf{z}_j\}_{j=1}^{N_2} \sim \bar{q}(\boldsymbol{\theta}). \quad (61)$$

Figure 1 summarizes the connections among the Eqs. (49), (59), (60) and the corresponding different methods. The standard IS and RIS schemes have been described in the previous sections, whereas the corresponding *locally-restricted* versions will be introduced below.

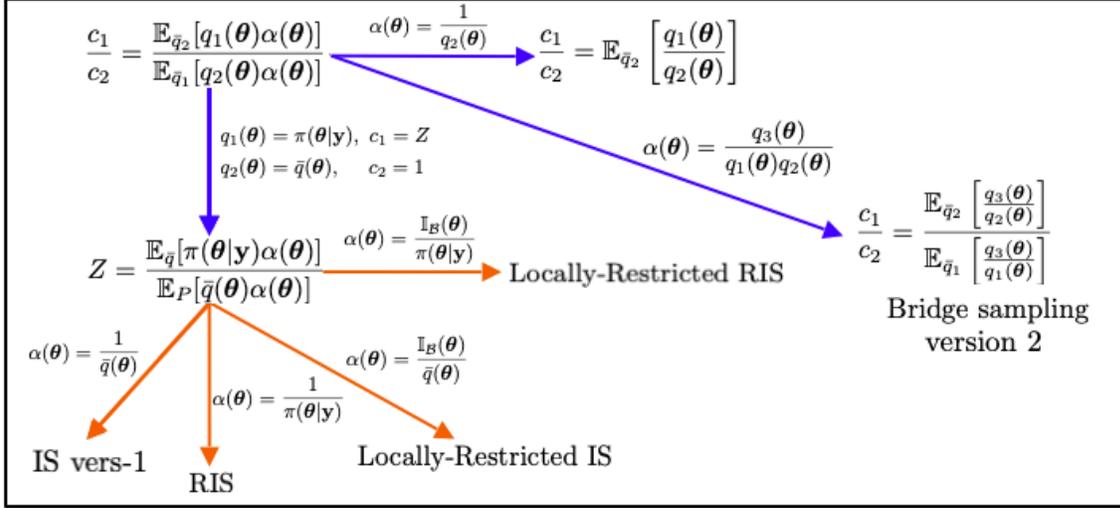


Figure 1: Graphical representation of the relationships among the Eqs. (49) (pre-umbrella identity), (59) (general bridge sampling identity), (60) (bridge sampling for Z) and the corresponding different methods, starting from bridge sampling identity (59).

4.2.2 Relationship with Chib's method

The Chib estimator, described in Section 3.4, is

$$\hat{Z} = \frac{\pi(\boldsymbol{\theta}^* | \mathbf{y}) \frac{1}{N_2} \sum_{j=1}^{N_2} \alpha(\boldsymbol{\theta}^*, \mathbf{v}_j)}{\frac{1}{N_1} \sum_{i=1}^{N_1} \alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*) \varphi(\boldsymbol{\theta}^* | \boldsymbol{\theta}_i)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim P(\boldsymbol{\theta} | \mathbf{y}), \quad \{\mathbf{v}_j\}_{j=1}^{N_2} \sim \varphi(\boldsymbol{\theta} | \boldsymbol{\theta}^*), \quad (62)$$

where $\varphi(\boldsymbol{\theta} | \boldsymbol{\theta}^*)$ is the proposal used inside an MCMC algorithm, $\alpha(\mathbf{x}, \mathbf{z}) : \mathbb{R}^{D_\theta} \times \mathbb{R}^{D_\theta} \rightarrow \mathbb{R}^+$ represents acceptance probability of this MCMC scheme and the point $\boldsymbol{\theta}^*$ is usually chosen in a high probability region. Note that the balance condition involving the function φ , P and α must be satisfied,

$$\alpha(\boldsymbol{\theta}, \mathbf{z}) \varphi(\mathbf{z} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{y}) = \alpha(\mathbf{z}, \boldsymbol{\theta}) \varphi(\boldsymbol{\theta} | \mathbf{z}) \pi(\mathbf{z} | \mathbf{y}).$$

Using the balance condition above, if we replace $\alpha(\boldsymbol{\theta}^*, \mathbf{v}_j) = \frac{\alpha(\mathbf{v}_j, \boldsymbol{\theta}^*)\varphi(\boldsymbol{\theta}^*|\mathbf{v}_j)\pi(\mathbf{v}_j|\mathbf{y})}{\varphi(\mathbf{v}_j|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*|\mathbf{y})}$ inside the numerator of (62), we obtain

$$\widehat{Z} = \frac{\frac{1}{N_2} \sum_{j=1}^{N_2} \frac{\varphi(\boldsymbol{\theta}^*|\mathbf{v}_j)}{\varphi(\mathbf{v}_j|\boldsymbol{\theta}^*)} \alpha(\mathbf{v}_j, \boldsymbol{\theta}^*) \pi(\mathbf{v}_j|\mathbf{y})}{\frac{1}{N_1} \sum_{i=1}^{N_1} \alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*) \varphi(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)}, \quad (63)$$

and if we also assume a symmetric proposal $\varphi(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = \varphi(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, we can finally write

$$\widehat{Z} = \frac{\frac{1}{N_2} \sum_{j=1}^{N_2} \alpha(\mathbf{v}_j, \boldsymbol{\theta}^*) \pi(\mathbf{v}_j|\mathbf{y})}{\frac{1}{N_1} \sum_{i=1}^{N_1} \alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*) \varphi(\boldsymbol{\theta}_i|\boldsymbol{\theta}^*)}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim P(\boldsymbol{\theta}|\mathbf{y}), \quad \{\mathbf{v}_j\}_{j=1}^{N_2} \sim \varphi(\boldsymbol{\theta}|\boldsymbol{\theta}^*), \quad (64)$$

We can observe a clear connection between the estimators (61) and (62). Clearly, $\varphi(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ plays the role of $\bar{q}(\boldsymbol{\theta})$ in (61), and the acceptance function $\alpha(\mathbf{x}, \mathbf{z})$ plays the role of the α function in (61). However, in this case, $\varphi(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ participates also *inside* the MCMC used for generating $\{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim P(\boldsymbol{\theta}|\mathbf{y})$. The function α takes also part to the generation MCMC chain, $\{\boldsymbol{\theta}_i\}_{i=1}^{N_1}$ (being the acceptance probability of the new states), and generally its evaluation involves the evaluation of φ and P . Note also that (62) is more generic than (64), being valid also for non-symmetric proposals φ . For further discussion see [36].

4.2.3 Locally-restricted IS and RIS

In the literature, there exist variants of the estimators in Eqs. (43) and (46). These corrected estimators are attempts to improve the efficiency (e.g., remove the infinite variance cases, specially in the harmonic estimator) by restricting the integration to a smaller subset of Θ (usually chosen in high posterior/likelihood-valued regions) generally denoted by $\mathcal{B} \subset \Theta$. As an example, \mathcal{B} can be a rectangular or ellipsoidal region centered at the *maximum a posteriori* (MAP) estimate $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$.

Locally-restricted IS estimator. Consider the posterior mass of subset $\mathcal{B} \subset \Theta$,

$$Z_{\mathcal{B}} = \int_{\mathcal{B}} P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int_{\Theta} \mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta}) \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{Z} d\boldsymbol{\theta}, \quad (65)$$

where $\mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta})$ is an indicator function, taking value 1 for $\boldsymbol{\theta} \in \mathcal{B}$ and 0 otherwise. It leads to the following representation

$$Z = \frac{1}{Z_{\mathcal{B}}} \int_{\Theta} \mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta}) \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{Z_{\mathcal{B}}} \mathbb{E}_{\bar{q}} \left[\mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta}) \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{\bar{q}(\boldsymbol{\theta})} \right]. \quad (66)$$

We can estimate $Z_{\mathcal{B}}$ considering N_1 samples from $P(\boldsymbol{\theta}|\mathbf{y})$ by taking the proportion of samples inside \mathcal{B} . The resulting locally-restricted IS estimator of Z is

$$\widehat{Z} = \frac{\frac{1}{N_1} \sum_{i=1}^{N_1} \frac{\mathbb{I}_{\mathcal{B}}(\mathbf{z}_i) \ell(\mathbf{y}|\mathbf{z}_i) g(\mathbf{z}_i)}{\bar{q}(\mathbf{z}_i)}}{\frac{1}{N_2} \sum_{i=1}^{N_2} \mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta}_i)}, \quad \{\mathbf{z}_i\}_{i=1}^{N_1} \sim \bar{q}(\boldsymbol{\theta}), \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_2} \sim P(\boldsymbol{\theta}|\mathbf{y}) \quad (\text{via MCMC}). \quad (67)$$

Note that the above estimator requires samples from two densities, namely the proposal $\bar{q}(\boldsymbol{\theta})$ and the posterior density $P(\boldsymbol{\theta}|\mathbf{y})$ (via MCMC).

Locally-restricted RIS estimator. To derive the locally-restricted RIS estimator, consider the mass of \mathcal{B} under $\bar{q}(\boldsymbol{\theta})$,

$$\bar{Q}(\mathcal{B}) = \int_{\mathcal{B}} \bar{q}(\boldsymbol{\theta}) d\boldsymbol{\theta} = Z \cdot \mathbb{E}_P \left[\mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta}) \frac{\bar{q}(\boldsymbol{\theta})}{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})} \right], \quad (68)$$

which leads to the following representation

$$Z = \frac{\bar{Q}(\mathcal{B})}{\mathbb{E}_P \left[\frac{\mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta})\bar{q}(\boldsymbol{\theta})}{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})} \right]}. \quad (69)$$

$\bar{Q}(\mathcal{B})$ can be estimated using a sample from $\bar{q}(\boldsymbol{\theta})$ by taking the proportion of sampled values inside \mathcal{B} . The locally-restricted RIS estimator is

$$\hat{Z} = \frac{\frac{1}{N_1} \sum_{i=1}^{N_1} \mathbb{I}_{\mathcal{B}}(\mathbf{z}_i)}{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta}_i)\bar{q}(\boldsymbol{\theta}_i)}{\ell(\mathbf{y}|\boldsymbol{\theta}_i)g(\boldsymbol{\theta}_i)}}, \quad \{\mathbf{z}_i\}_{i=1}^{N_1} \sim \bar{q}(\boldsymbol{\theta}), \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_2} \sim P(\boldsymbol{\theta}|\mathbf{y}). \quad (70)$$

Other variants, where \mathcal{B} corresponds to highest density regions, can be found in [46].

4.2.4 Optimal construction of bridge sampling

Identities as (59) are associated to the bridge sampling approach. However, considering $\alpha(\boldsymbol{\theta}) = \frac{q_3(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})q_1(\boldsymbol{\theta})}$ in Eq. (59), bridge sampling can be also motivated from the expression

$$\frac{c_1}{c_2} = \frac{c_3/c_2}{c_3/c_1} = \frac{\mathbb{E}_{\bar{q}_2} \left[\frac{q_3(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})} \right]}{\mathbb{E}_{\bar{q}_1} \left[\frac{q_3(\boldsymbol{\theta})}{q_1(\boldsymbol{\theta})} \right]}, \quad (71)$$

where the density $\bar{q}_3(\boldsymbol{\theta}) \propto q_3(\boldsymbol{\theta})$ is in some sense “in between” $q_1(\boldsymbol{\theta})$ and $q_2(\boldsymbol{\theta})$. That is, instead of applying directly (49) to $\frac{c_1}{c_2}$, we apply it to first estimate $\frac{c_3}{c_2}$ and $\frac{c_3}{c_1}$, and then take the ratio to cancel c_3 . The bridge sampling estimator of $\frac{c_1}{c_2}$ is then

$$\frac{c_1}{c_2} \approx \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{q_3(\mathbf{z}_i)}{q_2(\mathbf{z}_i)}}{\frac{1}{N_1} \sum_{i=1}^{N_1} \frac{q_3(\boldsymbol{\theta}_i)}{q_1(\boldsymbol{\theta}_i)}}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim \bar{q}_1(\boldsymbol{\theta}), \quad \{\mathbf{z}_i\}_{i=1}^{N_2} \sim \bar{q}_2(\boldsymbol{\theta}). \quad (72)$$

Remark 9. We do not need to draw samples from $\bar{q}_3(\boldsymbol{\theta})$, but only evaluate $q_3(\boldsymbol{\theta})$. For a comparison with umbrella sampling see Table 7.

Table 7: Joint use of three densities: comparison between bridge and umbrella sampling.

Method	$\bar{q}_1(\boldsymbol{\theta})$	$\bar{q}_3(\boldsymbol{\theta})$	$\bar{q}_2(\boldsymbol{\theta})$	Identity
Umbrella sampling	evaluate	draw from	evaluate	$\frac{c_1}{c_2} = \frac{c_1/c_3}{c_2/c_3} - (50)$
Bridge sampling	draw from	evaluate	draw from	$\frac{c_1}{c_2} = \frac{c_3/c_2}{c_3/c_1} - (71)$

Optimal bridge density. It can be shown that the optimal bridge density $\bar{q}_3(\boldsymbol{\theta})$ can be expressed as a weighted harmonic mean of $\bar{q}_1(\boldsymbol{\theta})$ and $\bar{q}_2(\boldsymbol{\theta})$ (with weights being the sampling rates),

$$\begin{aligned}
 \bar{q}_3^{\text{opt}}(\boldsymbol{\theta}) &= \frac{1}{\frac{N_2}{N_1+N_2} [\bar{q}_1(\boldsymbol{\theta})]^{-1} + \frac{N_1}{N_1+N_2} [\bar{q}_2(\boldsymbol{\theta})]^{-1}} \\
 &= \frac{1}{c_2} \cdot \frac{N_1 + N_2}{N_2 \frac{c_1}{c_2} q_1^{-1}(\boldsymbol{\theta}) + N_1 q_2^{-1}(\boldsymbol{\theta})} \\
 &\propto q_3^{\text{opt}}(\boldsymbol{\theta}) = \frac{q_1(\boldsymbol{\theta})q_2(\boldsymbol{\theta})}{N_1 q_1(\boldsymbol{\theta}) + N_2 \frac{c_1}{c_2} q_2(\boldsymbol{\theta})}. \tag{73}
 \end{aligned}$$

This is an optimal bridge density if both N_i are strictly positive, $N_i > 0$, hence we draw from both $\bar{q}_i(\boldsymbol{\theta})$. Note that $\bar{q}_3^{\text{opt}}(\boldsymbol{\theta})$ depends on the unknown ratio $r = \frac{c_1}{c_2}$. Therefore, we cannot even evaluate $q_3^{\text{opt}}(\boldsymbol{\theta})$. Hence, we need to resort to the following iterative procedure to approximate the optimal bridge sampling estimator. Noting that

$$\frac{q_3^{\text{opt}}(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})} = \frac{q_1(\boldsymbol{\theta})}{N_1 q_1(\boldsymbol{\theta}) + r N_2 q_2(\boldsymbol{\theta})}, \quad \frac{q_3^{\text{opt}}(\boldsymbol{\theta})}{q_1(\boldsymbol{\theta})} = \frac{q_2(\boldsymbol{\theta})}{N_1 q_1(\boldsymbol{\theta}) + r N_2 q_2(\boldsymbol{\theta})}. \tag{74}$$

The iterative procedure is formed by the following steps:

1. Start with an initial estimate $\hat{r}^{(1)} \approx \frac{c_1}{c_2}$ (using e.g. Laplace's).
2. For $t = 1, \dots, T$:

- (a) Draw $\{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim \bar{q}_1(\boldsymbol{\theta})$ and $\{\mathbf{z}_i\}_{i=1}^{N_2} \sim \bar{q}_2(\boldsymbol{\theta})$ and iterate

$$\hat{r}^{(t+1)} = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{q_1(\mathbf{z}_i)}{N_1 q_1(\mathbf{z}_i) + N_2 \hat{r}^{(t)} q_2(\mathbf{z}_i)}}{\frac{1}{N_1} \sum_{i=1}^{N_1} \frac{q_2(\boldsymbol{\theta}_i)}{N_1 q_1(\boldsymbol{\theta}_i) + N_2 \hat{r}^{(t)} q_2(\boldsymbol{\theta}_i)}}. \tag{75}$$

Remark 10. In [48, Theorem 3.3], the authors show that the asymptotic error of optimal bridge sampling with \bar{q}_3^{opt} in Eq. (73) is always greater than the asymptotic error of optimal umbrella sampling using $\bar{q}_3^{\text{opt}}(\boldsymbol{\theta}) \propto |\bar{q}_1(\boldsymbol{\theta}) - \bar{q}_2(\boldsymbol{\theta})|$ in Eq. (51).

Optimal bridge sampling for Z . Given the considerations above, an iterative bridge sampling

estimator of Z is obtained by setting $q_1(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$, $c_1 = Z$, $\bar{q}_2(\boldsymbol{\theta}) = \bar{q}(\boldsymbol{\theta})$, so that

$$\widehat{Z}^{(t+1)} = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\pi(\mathbf{z}_i|\mathbf{y})}{N_1\pi(\mathbf{z}_i|\mathbf{y}) + N_2Z^{(t)}\bar{q}(\mathbf{z}_i)}}{\frac{1}{N_1} \sum_{i=1}^{N_1} \frac{\bar{q}(\boldsymbol{\theta}_i)}{N_1\pi(\boldsymbol{\theta}_i|\mathbf{y}) + N_2Z^{(t)}\bar{q}(\boldsymbol{\theta}_i)}}, \quad \{\mathbf{z}_i\}_{i=1}^{N_2} \sim \bar{q}(\boldsymbol{\theta}) \text{ and } \{\boldsymbol{\theta}_i\}_{i=1}^{N_1} \sim P(\boldsymbol{\theta}|\mathbf{y}). \quad (76)$$

for $t = 1, \dots, T$. Looking at Eqs. (73) and (71), when $N_1 = 0$, that is, when all samples are drawn from $\bar{q}(\boldsymbol{\theta})$, the estimator above reduces to (non-iterative) standard IS scheme with proposal $\bar{q}(\boldsymbol{\theta})$. When $N_2 = 0$, that is, when all samples are drawn from $P(\boldsymbol{\theta}|\mathbf{y})$, the estimator becomes the (non-iterative) RIS estimator. See [48] for a comparison of optimal umbrella sampling, bridge sampling and path sampling (described in the next section). An alternative derivation of the optimal bridge sampling estimator is given in [46], by generating samples from a mixture of type $\psi(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}|\mathbf{y}) + v\bar{q}(\boldsymbol{\theta})$. However, the resulting estimator employs the same samples drawn from $\psi(\boldsymbol{\theta})$ in the numerator and denominator, unlike in Eq. (76).

4.2.5 Other estimators drawing from a generic proposal and the posterior

Let consider again the scenario where we have a set of samples $\{\boldsymbol{\theta}_i\}_{i=1}^{N_1}$ from the posterior $P(\boldsymbol{\theta}|\mathbf{y})$ and set $\{\mathbf{z}_i\}_{i=1}^{N_2}$ from some proposal $\bar{q}(\boldsymbol{\theta})$, as in the bridge sampling case described above. However, here we consider that these two sets $\{\tilde{\boldsymbol{\theta}}_i\}_{i=1}^{N_1+N_2} = \{\{\boldsymbol{\theta}_i\}_{i=1}^{N_1}, \{\mathbf{z}_i\}_{i=1}^{N_2}\}$ are drawn from the mixture $\bar{q}_{\text{mix}}(\boldsymbol{\theta}) = \frac{N_1}{N_1+N_2}P(\boldsymbol{\theta}|\mathbf{y}) + \frac{N_2}{N_1+N_2}\bar{q}(\boldsymbol{\theta})$ considering a deterministic mixture sampling approach [51]. Thus, we can use the IS identities that use a *single* proposal, namely Eqs. (49) and (50).

Importance sampling with mixture (M-IS). Setting $\bar{q}_1(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$ and $\bar{q}_2(\boldsymbol{\theta}) = \bar{q}_{\text{mix}}(\boldsymbol{\theta})$ in Eq. (49), we have

$$\widehat{Z}_{\text{M-IS}} = \frac{1}{N_1 + N_2} \sum_{i=1}^{N_1+N_2} \frac{\pi(\tilde{\boldsymbol{\theta}}_i|\mathbf{y})}{\bar{q}_{\text{mix}}(\tilde{\boldsymbol{\theta}}_i)} = \frac{1}{N_1 + N_2} \sum_{i=1}^{N_1+N_2} \frac{\pi(\tilde{\boldsymbol{\theta}}_i|\mathbf{y})}{\frac{N_1}{N_1+N_2}P(\tilde{\boldsymbol{\theta}}_i|\mathbf{y}) + \frac{N_2}{N_1+N_2}\bar{q}(\tilde{\boldsymbol{\theta}}_i)}, \quad (77)$$

where $\tilde{\boldsymbol{\theta}}_i \sim \bar{q}_{\text{mix}}(\boldsymbol{\theta}) = \frac{N_1}{N_1+N_2}P(\boldsymbol{\theta}|\mathbf{y}) + \frac{N_2}{N_1+N_2}\bar{q}(\boldsymbol{\theta})$ [51]. This estimator cannot be directly used since it requires the evaluation of $P(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{Z}\pi(\boldsymbol{\theta}|\mathbf{y})$. From an initial guess $\widehat{Z}^{(0)}$, the following iterative procedure can be used

$$\widehat{Z}^{(t)} = \frac{1}{N_1 + N_2} \sum_{i=1}^{N_1+N_2} \frac{\widehat{Z}^{(t-1)}\pi(\tilde{\boldsymbol{\theta}}_i|\mathbf{y})}{\frac{N_1}{N_1+N_2}\pi(\tilde{\boldsymbol{\theta}}_i|\mathbf{y}) + \frac{N_2}{N_1+N_2}\widehat{Z}^{(t-1)}\bar{q}(\tilde{\boldsymbol{\theta}}_i)}, \quad t \in \mathbb{N}. \quad (78)$$

Self-IS with mixture proposal (M-Self-IS). Setting $\bar{q}_1(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$, $\bar{q}_2(\boldsymbol{\theta}) = \bar{q}(\boldsymbol{\theta})$ and $\bar{q}_3(\boldsymbol{\theta}) = \bar{q}_{\text{mix}}$ in Eq. (50), we have

$$\widehat{Z}_{\text{M-Self-IS}} = \frac{\sum_{i=1}^{N_1+N_2} \frac{\pi(\tilde{\boldsymbol{\theta}}_i|\mathbf{y})}{\bar{q}_{\text{mix}}(\tilde{\boldsymbol{\theta}}_i)}}{\sum_{i=1}^{N_1+N_2} \frac{\bar{q}(\tilde{\boldsymbol{\theta}}_i)}{\bar{q}_{\text{mix}}(\tilde{\boldsymbol{\theta}}_i)}} = \frac{\sum_{i=1}^{N_1+N_2} \frac{\pi(\tilde{\boldsymbol{\theta}}_i|\mathbf{y})}{\frac{N_1}{N_1+N_2}P(\tilde{\boldsymbol{\theta}}_i|\mathbf{y}) + \frac{N_2}{N_1+N_2}\bar{q}(\tilde{\boldsymbol{\theta}}_i)}}{\sum_{i=1}^{N_1+N_2} \frac{\bar{q}(\tilde{\boldsymbol{\theta}}_i)}{\frac{N_1}{N_1+N_2}P(\tilde{\boldsymbol{\theta}}_i|\mathbf{y}) + \frac{N_2}{N_1+N_2}\bar{q}(\tilde{\boldsymbol{\theta}}_i)}}, \quad (79)$$

where $\tilde{\theta}_i \sim \bar{q}_{\text{mix}} = \frac{N_1}{N_1+N_2}P(\boldsymbol{\theta}|\mathbf{y}) + \frac{N_2}{N_1+N_2}\bar{q}(\boldsymbol{\theta})$ (drawn in a deterministic way). As above, this estimator is not of direct use, so we need to iterate

$$\widehat{Z}^{(t)} = \frac{\sum_{i=1}^{N_1+N_2} \frac{\pi(\tilde{\theta}_i|\mathbf{y})}{\frac{N_1}{N_1+N_2}\pi(\tilde{\theta}_i) + \frac{N_2}{N_1+N_2}\widehat{Z}^{(t-1)}\bar{q}(\tilde{\theta}_i)}}{\sum_{i=1}^{N_1+N_2} \frac{\bar{q}(\tilde{\theta}_i)}{\frac{N_1}{N_1+N_2}\pi(\tilde{\theta}_i|\mathbf{y}) + \frac{N_2}{N_1+N_2}\widehat{Z}^{(t-1)}\bar{q}(\tilde{\theta}_i)}}, \quad t \in \mathbb{N}. \quad (80)$$

This iterative estimator is very similar to the iterative optimal bridge sampling estimator in Eq. (76), but it uses both set of samples in numerator and denominator. This estimator is also related to the reverse logistic regression method in [53] (for more details see [48, 54], and the next section). Furthermore, the iterative estimator (80) is also discussed for the case $\bar{q}(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ in [55], in an attempt to exploit the advantages of the Naive Monte Carlo and the harmonic mean estimators, while removing their drawbacks.

Remark 11. *Both iterative versions (78)-(80) converge to the optimal bridge sampling estimator (76). See [52], for a related discussion. As we show in the simulation study, the speed of convergence of each iterative method is different. The iterative bridge sampling estimator seems to be the quickest one.*

4.2.6 Summary

Several techniques described in the last two subsections, including both umbrella and bridge sampling, are encompassed by the generic formula

$$\frac{c_1}{c_2} = \mathbb{E}_{\xi}[q_1(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})] / \mathbb{E}_{\bar{\chi}}[q_2(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})] \quad (81)$$

as shown in Table 8. The techniques differ also for which densities are drawn from and which densities are just evaluated.

4.3 IS based on multiple proposal densities

In this section we consider estimators of Z using samples drawn from more than two proposal densities. These schemes are usually based on the so-called tempering and/or annealing approach.

Reasons for tempering. The idea is again to consider densities that are in some sense “in the middle” between the posterior $P(\boldsymbol{\theta}|\mathbf{y})$ and an easier-to-work-with density (e.g. the prior $g(\boldsymbol{\theta})$ or some other proposal density). These densities are usually scaled version of the posterior. Generally, the scale parameter is called *temperature*.³ For this reason, the resulting pdfs are usually named tempered posteriors and correspond to flatter, more diffuse distributions than the standard posterior. The use of the tempered pdfs usually improve the mixing of the MCMC algorithms and foster the exploration of the space Θ . Generally, it helps the Monte Carlo methods (as MCMC and IS) to find the regions of posterior high probability. The number of such middle densities is

³The data tempering is also possible: the tempered posteriors contain less data than the complete posterior.

Table 8: Summary of the IS schemes (with one or two proposal pdfs), using Eq. (81).

$\frac{c_1}{c_2} = \mathbb{E}_{\bar{\xi}}[q_1(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})] / \mathbb{E}_{\bar{\chi}}[q_2(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})]$								
<i>For estimating a generic ratio c_1/c_2</i>								
Name	$\alpha(\boldsymbol{\theta})$	$\xi(\boldsymbol{\theta})$	$\bar{\chi}(\boldsymbol{\theta})$	$q_1(\boldsymbol{\theta})$	$q_2(\boldsymbol{\theta})$	c_1	c_2	sampling from
Bridge Identity - Eq. (59)	$\alpha(\boldsymbol{\theta})$	$\bar{q}_2(\boldsymbol{\theta})$	$\bar{q}_1(\boldsymbol{\theta})$					$\bar{q}_1(\boldsymbol{\theta}), \bar{q}_2(\boldsymbol{\theta})$
Bridge Identity - Eq. (71)	$\frac{q_3(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})q_1(\boldsymbol{\theta})}$	$\bar{q}_2(\boldsymbol{\theta})$	$\bar{q}_1(\boldsymbol{\theta})$	$q_1(\boldsymbol{\theta})$	$q_2(\boldsymbol{\theta})$	c_1	c_2	$\bar{q}_1(\boldsymbol{\theta}), \bar{q}_2(\boldsymbol{\theta})$
Identity - Eq. (49)	$\frac{1}{q_2(\boldsymbol{\theta})}$	$\bar{q}_2(\boldsymbol{\theta})$	$\bar{q}_1(\boldsymbol{\theta})$					$\bar{q}_2(\boldsymbol{\theta})$
Umbrella - Eq. (50)	$\frac{1}{q_3(\boldsymbol{\theta})}$	$\bar{q}_3(\boldsymbol{\theta})$	$\bar{q}_3(\boldsymbol{\theta})$					$\bar{q}_3(\boldsymbol{\theta})$
<i>For estimating Z, with one proposal</i>								
Self-norm. IS - Eq. (55)	$\frac{1}{q_3(\boldsymbol{\theta})}$	$\bar{q}_3(\boldsymbol{\theta})$	$\bar{q}_3(\boldsymbol{\theta})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	$f(\boldsymbol{\theta})$			$\bar{q}_3(\boldsymbol{\theta})$
IS vers-1	$1/\bar{q}(\boldsymbol{\theta})$	$\bar{q}(\boldsymbol{\theta})$	$P(\boldsymbol{\theta} \mathbf{y})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$	Z	1	$\bar{q}(\boldsymbol{\theta})$
RIS	$1/\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$	$P(\boldsymbol{\theta} \mathbf{y})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$			$P(\boldsymbol{\theta} \mathbf{y})$
<i>For estimating Z, with two proposals, $P(\boldsymbol{\theta} \mathbf{y})$ and $\bar{q}(\boldsymbol{\theta})$</i>								
Bridge Identity - Eq. (60)	$\alpha(\boldsymbol{\theta})$	$\bar{q}(\boldsymbol{\theta})$	$P(\boldsymbol{\theta} \mathbf{y})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$			
Locally-Restricted IS	$\mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta})/\bar{q}(\boldsymbol{\theta})$	$\bar{q}(\boldsymbol{\theta})$	$P(\boldsymbol{\theta} \mathbf{y})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$	Z	1	$P(\boldsymbol{\theta} \mathbf{y}), \bar{q}(\boldsymbol{\theta})$
Locally-Restricted RIS	$\mathbb{I}_{\mathcal{B}}(\boldsymbol{\theta})/\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$	$P(\boldsymbol{\theta} \mathbf{y})$	$\pi(\boldsymbol{\theta} \mathbf{y})$	$\bar{q}(\boldsymbol{\theta})$			

specified by the user, and in some cases, it is equivalent to the selection of a temperature schedule for linking the prior $g(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\mathbf{y})$. This idea is shared by the several methods, such as path sampling, power posterior methods and stepping-stone sampling described below.

First of all, we start with a general IS scheme considering different proposals $\bar{q}_n(\boldsymbol{\theta})$'s. Some of them could be tempered posteriors and the generation would be performed by an MCMC method in this case.

4.3.1 Multiple Importance Sampling (MIS) estimators

Here, we consider to generate samples from different proposal densities, i.e.,

$$\boldsymbol{\theta}_n \sim \bar{q}_n(\boldsymbol{\theta}), \quad n = 1, \dots, N. \quad (82)$$

In this scenario, different proper importance weights can be used [51, 56, 57]. The most efficient MIS scheme considers the following weights

$$w_n = \frac{\pi(\boldsymbol{\theta}_n|\mathbf{y})}{\frac{1}{N} \sum_{i=1}^N \bar{q}_i(\boldsymbol{\theta}_n)} = \frac{\pi(\boldsymbol{\theta}_n|\mathbf{y})}{\psi(\boldsymbol{\theta}_n)}, \quad (83)$$

where $\psi(\boldsymbol{\theta}_n) = \frac{1}{N} \sum_{i=1}^N \bar{q}_i(\boldsymbol{\theta}_n)$. Indeed, considering the set of samples $\{\boldsymbol{\theta}_n\}_{n=1}^N$ drawn in a deterministic order, $\boldsymbol{\theta}_n \sim \bar{q}_n(\boldsymbol{\theta})$, and given a sample $\boldsymbol{\theta}^* \in \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N\}$ uniformly chosen in $\{\boldsymbol{\theta}_n\}_{n=1}^N$,

then we can write $\boldsymbol{\theta}^* \sim \psi(\boldsymbol{\theta}_n)$. The standard MIS estimator is

$$\widehat{Z} = \frac{1}{N} \sum_{n=1}^N w_n = \frac{1}{N} \sum_{n=1}^N \frac{\pi(\boldsymbol{\theta}_n|\mathbf{y})}{\psi(\boldsymbol{\theta}_n)} \quad (84)$$

$$= \frac{1}{N} \sum_{n=1}^N \frac{g(\boldsymbol{\theta}_n)\ell(\mathbf{y}|\boldsymbol{\theta}_n)}{\psi(\boldsymbol{\theta}_n)}, \quad (85)$$

$$= \frac{1}{N} \sum_{n=1}^N \eta_n \ell(\mathbf{y}|\boldsymbol{\theta}_n), \quad \boldsymbol{\theta}_n \sim \bar{q}_n(\boldsymbol{\theta}), \quad n = 1, \dots, N. \quad (86)$$

where $\eta_n = \frac{g(\boldsymbol{\theta}_n)}{\psi(\boldsymbol{\theta}_n)}$. The estimator is unbiased [51]. As in the standard IS scheme, an alternative biased estimator is

$$\widehat{Z} = \sum_{n=1}^N \bar{\eta}_n \ell(\mathbf{y}|\boldsymbol{\theta}_n), \quad \boldsymbol{\theta}_n \sim \bar{q}_n(\boldsymbol{\theta}), \quad n = 1, \dots, N, \quad (87)$$

where $\bar{\eta}_n = \frac{\eta_n}{\sum_{i=1}^N \eta_i}$, so that $\sum_{i=1}^N \bar{\eta}_i = 1$ and we have a convex combination of likelihood values $\ell(\mathbf{y}|\boldsymbol{\theta}_n)$'s. It is a generalization of the estimator in Eq. (40) and recalled below in Eq. (88).

4.3.2 Tempered posteriors as proposal densities

Let recall the IS vers-2 estimator of Z in Eq. (40), which involves a weighted sum of likelihood evaluations at points $\{\boldsymbol{\theta}_i\}_{i=1}^N$ drawn from importance density $\bar{q}(\boldsymbol{\theta})$ (but we can evaluate only $q(\boldsymbol{\theta}) \propto \bar{q}(\boldsymbol{\theta})$),

$$\widehat{Z} = \sum_{i=1}^N \bar{\rho}_i \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad \bar{\rho}_i = \frac{\frac{g(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}}{\sum_{n=1}^N \frac{g(\boldsymbol{\theta}_n)}{q(\boldsymbol{\theta}_n)}} \propto \frac{g(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}, \quad (88)$$

where $\sum_{i=1}^N \bar{\rho}_i = 1$. Let us consider

$$\bar{q}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y}, \beta) \propto q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y}, \beta) = g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^\beta,$$

with $\beta \in [0, 1]$. Namely, we use a tempered posterior as importance density. Note that we can evaluate only the unnormalized density $q(\boldsymbol{\theta})$. The IS estimator version 2 can be employed in this case, and we obtain $\bar{\rho}_i \propto \frac{g(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)\ell(\mathbf{y}|\boldsymbol{\theta}_i)^\beta} = \frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta}_i)^\beta}$. The resulting IS estimator version 2 is

$$\widehat{Z} = \frac{\sum_{i=1}^N \frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta}_i)^\beta} \ell(\mathbf{y}|\boldsymbol{\theta}_i)}{\sum_{i=1}^N \frac{1}{\ell(\mathbf{y}|\boldsymbol{\theta}_i)^\beta}} \quad (89)$$

$$= \frac{\sum_{i=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_i)^{1-\beta}}{\sum_{i=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_i)^{-\beta}} \quad \{\boldsymbol{\theta}_i\}_{i=1}^N \sim P(\boldsymbol{\theta}|\mathbf{y}, \beta) \quad (\text{via MCMC}). \quad (90)$$

This method is denoted below as IS with a tempered posterior as proposal (IS-P). Table 9 shows that this technique includes different schemes for different values of β . Different possible MIS schemes can be also considered, i.e., using Eq. (87) for instance [51, 57].

Table 9: Different estimators of Z using $\bar{q}(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^\beta$ as importance density, with $\beta \in [0, 1]$.

Name	Coefficient β	Weights $\bar{\rho}_i$	Estimator $\widehat{Z} = \sum_{i=1}^N \bar{\rho}_i \ell(\mathbf{y} \boldsymbol{\theta}_i)$
Naive Monte Carlo	$\beta = 0$	$\frac{1}{N}$	$\frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y} \boldsymbol{\theta}_i)$
Harmonic Mean Estimator	$\beta = 1$	$\frac{\frac{1}{\ell(\mathbf{y} \boldsymbol{\theta}_i)}}{\sum_j \frac{1}{\ell(\mathbf{y} \boldsymbol{\theta}_j)}}$	$\widehat{Z} = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{1}{\ell(\mathbf{y} \boldsymbol{\theta}_i)}}$
Power posterior as proposal pdf	$0 < \beta < 1$	$\frac{\frac{1}{\ell(\mathbf{y} \boldsymbol{\theta}_i)^\beta}}{\sum_j \frac{1}{\ell(\mathbf{y} \boldsymbol{\theta}_j)^\beta}}$	$\widehat{Z} = \frac{\sum_i \ell(\mathbf{y} \boldsymbol{\theta}_i)^{1-\beta}}{\sum_i \ell(\mathbf{y} \boldsymbol{\theta}_i)^{-\beta}}$

Remark 12. One could consider also to draw samples from N different tempered posteriors, $\boldsymbol{\theta}_n \sim P(\boldsymbol{\theta}|\mathbf{y}, \beta_n) \propto g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^{\beta_n}$, with $n = 1, \dots, N$, and then apply deterministic mixture idea in (87). However, in this case, we cannot evaluate properly the mixture

$$\psi(\boldsymbol{\theta}_n) = \frac{1}{N} \sum_{i=1}^N P(\boldsymbol{\theta}_n|\mathbf{y}, \beta_i) = \frac{1}{N} \sum_{i=1}^N \frac{1}{Z(\beta_i)} \pi(\boldsymbol{\theta}_n|\mathbf{y}, \beta_i).$$

Here, the issue is a not just a global unknown normalizing constant (as usual): in this case, we do not know the weights of the mixture since all $Z(\beta) = \int_{\Theta} g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^\beta d\boldsymbol{\theta}$ are unknown. This problem can be solved using the techniques described in the next sections.

Reverse logistic regression (RLR). In RLR, the idea is to apply IS with the mixture $\psi(\boldsymbol{\theta}_n)$ in the remark above. The normalizing constants $Z(\beta_i)$ are iteratively obtained by maximizing of a suitable log-likelihood, built with the samples from each tempered posterior $P(\boldsymbol{\theta}|\mathbf{y}, \beta_n)$ [53, 58, 54].

In the next section, we describe an alternative to RLR for employing different tempered posteriors as proposals.

4.3.3 Stepping-stone (SS) sampling

Consider again $P(\boldsymbol{\theta}|\mathbf{y}, \beta) \propto g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^\beta$ and $Z(\beta) = \int_{\Theta} g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^\beta d\boldsymbol{\theta}$. The goal is to estimate $Z = \frac{Z(1)}{Z(0)}$, which can be expressed as the following product, with $\beta_0 = 0$ and $\beta_K = 1$,

$$Z = \frac{Z(1)}{Z(0)} = \prod_{k=1}^K \frac{Z(\beta_k)}{Z(\beta_{k-1})}, \quad (91)$$

where β_k are often chosen as $\beta_k = \frac{k}{K}$, $k = 1, \dots, K$, i.e., with a uniform grid in $[0, 1]$. Note that generally $Z(0) = 1$, since it is normalizing constant of the prior. The SS method is based on the following identity,

$$\begin{aligned} \mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})} \left[\frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_k)}{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})} \right] &= \int_{\Theta} \frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_k)}{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})} P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1}) d\boldsymbol{\theta}, \\ &= \frac{1}{Z(\beta_{k-1})} \int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y}, \beta_k) d\boldsymbol{\theta} = \frac{Z(\beta_k)}{Z(\beta_{k-1})}. \end{aligned}$$

Then, the idea of SS sampling is to estimate each ratio $r_k = \frac{Z(\beta_k)}{Z(\beta_{k-1})}$ by importance sampling as

$$r_k = \frac{Z(\beta_k)}{Z(\beta_{k-1})} = \mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})} \left[\frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_k)}{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})} \right] \quad (92)$$

$$= \mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})} \left[\frac{\ell(\mathbf{y}|\boldsymbol{\theta})^{\beta_k}}{\ell(\mathbf{y}|\boldsymbol{\theta})^{\beta_{k-1}}} \right] \quad (93)$$

$$\approx \hat{r}_k = \frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_{i,k-1})^{\beta_k - \beta_{k-1}}, \quad \{\boldsymbol{\theta}_{i,k-1}\}_{i=1}^N \sim P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1}). \quad (94)$$

Multiplying all ratio estimates yields the final estimator of Z

$$\hat{Z} = \prod_{k=1}^K \hat{r}_k = \prod_{k=1}^K \left(\frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_{i,k-1})^{\beta_k - \beta_{k-1}} \right), \quad \{\boldsymbol{\theta}_{i,k-1}\}_{i=1}^N \sim P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1}). \quad (95)$$

For $K = 1$, we come back to the Naive MC estimator. The sampling procedure of the SS method is graphically represented in Figure 2.

Remark 13. *The SS estimator is unbiased, since it is a product of unbiased estimators.*

The two following methods, path sampling and power posteriors, estimate $\log Z$ instead of Z .

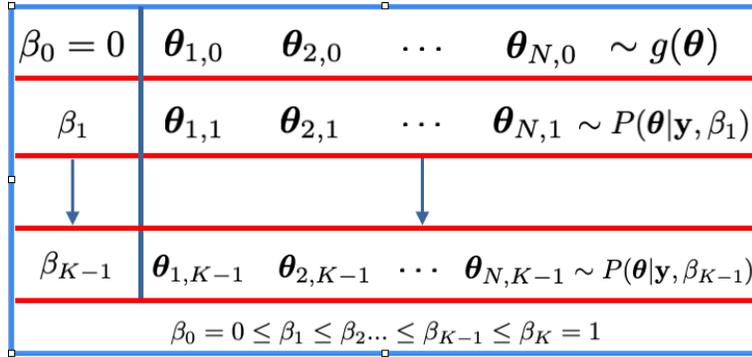


Figure 2: Sampling procedure in the SS method. Note that samples from $P(\boldsymbol{\theta}|\mathbf{y})$ ($\beta_K = 1$) are not considered. It is relevant to compare this figure with Figures 4-5 in the next section.

4.3.4 Path sampling (a.k.a., thermodynamic integration)

More specifically, the method of path sampling for estimating $\frac{c_1}{c_2}$ relies on the idea of building and drawing samples from a sequence of distributions linking $\bar{q}_1(\boldsymbol{\theta})$ and $\bar{q}_2(\boldsymbol{\theta})$ (a continuous path). For the purpose of estimating only one constant, the marginal likelihood Z , we set $\bar{q}_2(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ and $\bar{q}_1(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$ and we link them by a univariate path with parameter β . Let

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \beta), \quad \beta \in [0, 1], \quad (96)$$

denote a sequence of (probably unnormalized except for $\beta = 0$) densities such $\pi(\boldsymbol{\theta}|\mathbf{y}, \beta = 0) = g(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{y}, \beta = 1) = \pi(\boldsymbol{\theta}|\mathbf{y})$. More generally, we could consider $\pi(\boldsymbol{\theta}|\mathbf{y}, \beta = 0) = \bar{q}(\boldsymbol{\theta})$ where $\bar{q}(\boldsymbol{\theta})$ is a generic normalized proposal density, possibly closer to the posterior than $g(\boldsymbol{\theta})$. The path sampling method for estimating the marginal likelihood is based on expressing $\log Z$ as

$$\log Z = \mathbb{E}_{p(\boldsymbol{\theta}, \beta|\mathbf{y})} \left[\frac{U(\boldsymbol{\theta}, \beta)}{p(\beta)} \right], \quad \text{with } U(\boldsymbol{\theta}, \beta) = \frac{\partial}{\partial \beta} \log \pi(\boldsymbol{\theta}|\mathbf{y}, \beta), \quad (97)$$

where the expectation is w.r.t. the joint $p(\boldsymbol{\theta}, \beta|\mathbf{y}) = \frac{1}{Z(\beta)}\pi(\boldsymbol{\theta}|\mathbf{y}, \beta)p(\beta)$, being $Z(\beta)$ the normalizing constant of $\pi(\boldsymbol{\theta}|\mathbf{y}, \beta)$ and $p(\beta)$ represents a density for $\beta \in [0, 1]$. Indeed, we have

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{\theta}, \beta|\mathbf{y})} \left[\frac{U(\boldsymbol{\theta}, \beta)}{p(\beta)} \right] &= \int_{\Theta} \int_0^1 \frac{1}{p(\beta)} \left[\frac{\partial}{\partial \beta} \log \pi(\boldsymbol{\theta}|\mathbf{y}, \beta) \right] \frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta)}{Z(\beta)} p(\beta) d\boldsymbol{\theta} d\beta, \\ &= \int_{\Theta} \int_0^1 \frac{1}{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta)} \left[\frac{\partial}{\partial \beta} \pi(\boldsymbol{\theta}|\mathbf{y}, \beta) \right] \frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta)}{Z(\beta)} d\boldsymbol{\theta} d\beta, \\ &= \int_{\Theta} \int_0^1 \frac{1}{Z(\beta)} \frac{\partial}{\partial \beta} \pi(\boldsymbol{\theta}|\mathbf{y}, \beta) d\boldsymbol{\theta} d\beta, \\ &= \int_0^1 \frac{1}{Z(\beta)} \frac{\partial}{\partial \beta} \left(\int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{y}, \beta) d\boldsymbol{\theta} \right) d\beta, \\ &= \int_0^1 \frac{1}{Z(\beta)} \frac{\partial}{\partial \beta} Z(\beta) d\beta \\ &= \int_0^1 \frac{\partial}{\partial \beta} \log Z(\beta) d\beta = \log Z(1) - \log Z(0) = \log Z, \end{aligned} \quad (98)$$

where we substituted $Z(\beta = 1) = Z(1) = Z$ and $Z(\beta = 0) = Z(0) = 1$. Thus, using a sample $\{\boldsymbol{\theta}_i, \beta_i\}_{i=1}^N \sim p(\boldsymbol{\theta}, \beta|\mathbf{y})$, we can write the path sampling estimator for $\log Z$

$$\widehat{\log Z} = \frac{1}{N} \sum_{i=1}^N \frac{U(\boldsymbol{\theta}_i, \beta_i)}{p(\beta_i)}, \quad \{\boldsymbol{\theta}_i, \beta_i\}_{i=1}^N \sim p(\boldsymbol{\theta}, \beta|\mathbf{y}). \quad (99)$$

The samples from $p(\boldsymbol{\theta}, \beta|\mathbf{y})$ may be obtained by first drawing $\beta'(\beta_i)$ from $p(\beta)$ and then applying some MCMC steps to draw from $P(\boldsymbol{\theta}|\mathbf{y}, \beta') \propto \pi(\boldsymbol{\theta}|\mathbf{y}, \beta')$ given β' . Therefore, in path sampling, we have to choose (a) the path and (b) and the prior $p(\beta)$. A discussion regarding the optimal choices of the path and $p(\beta)$, see [59]. The optimal path for linking any two given densities is impractical as it depends on the normalizing constants being estimated. The geometric path described below, although suboptimal, is generic and simple to implement.

Geometric path. Often a geometric path is employed,

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{y}, \beta) &= g(\boldsymbol{\theta})^{1-\beta} \pi(\boldsymbol{\theta}|\mathbf{y})^\beta \\ &= g(\boldsymbol{\theta}) \ell(\mathbf{y}|\boldsymbol{\theta})^\beta, \quad \beta \in [0, 1]. \end{aligned} \quad (100)$$

Note that $\pi(\boldsymbol{\theta}|\mathbf{y}, \beta)$ is the posterior with a powered, “less informative” - “wider” likelihood (for this reason, $\pi(\boldsymbol{\theta}|\mathbf{y}, \beta)$ is often called a “power posterior”). In this case, we have

$$U(\boldsymbol{\theta}, \beta) = \frac{\partial}{\partial \beta} \log \pi(\boldsymbol{\theta}|\mathbf{y}, \beta) = \log \ell(\mathbf{y}|\boldsymbol{\theta}),$$

so the path sampling identity becomes

$$\log Z = \mathbb{E}_{p(\boldsymbol{\theta}, \beta|\mathbf{y})} \left[\frac{\log \ell(\mathbf{y}|\boldsymbol{\theta})}{p(\beta)} \right], \quad (101)$$

which is also used in the power posterior method of [60], described in Section 4.3.6.

4.3.5 Connections among path sampling, bridge sampling and stepping-stones

The path sampling method can be motivated from bridge sampling by applying the bridge sampling identity in (71) in a chain fashion. Assume we have $K + 1$ densities $P(\boldsymbol{\theta}|\mathbf{y}, \beta_k) = \pi(\boldsymbol{\theta}|\mathbf{y}, \beta_k)/Z(\beta_k)$, $k = 0, \dots, K$ from which we can draw samples, with endpoints $P(\boldsymbol{\theta}|\mathbf{y}, \beta_0 = 0) = g(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\mathbf{y}, \beta_K = 1) = P(\boldsymbol{\theta}|\mathbf{y})$. We can express $Z = Z(\beta_K) = Z(1)$ as follows

$$Z = \prod_{k=1}^K \frac{Z(\beta_k)}{Z(\beta_{k-1})} = \prod_{k=1}^K \frac{\mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})} \left[\frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-\frac{1}{2}})}{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_k)} \right]}{\mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y}, \beta_k)} \left[\frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-\frac{1}{2}})}{\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_k)} \right]}. \quad (102)$$

Note that we have applied the bridge sampling identity in Eq. (71) to each ratio $\frac{Z(\beta_k)}{Z(\beta_{k-1})}$, using $K - 1$ middle densities $\pi(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-\frac{1}{2}})$. We can approximate the k -th term by using samples from $P(\boldsymbol{\theta}|\mathbf{y}, \beta_{k-1})$ and $P(\boldsymbol{\theta}|\mathbf{y}, \beta_k)$, and take the product to obtain the final estimator of Z . Taking the logarithm of the above expression, as $K \rightarrow \infty$, results in the basic identity of path sampling for estimating Z in Eq. (97) [59]. In this sense, path sampling can be interpreted as a continuous application of bridge sampling steps. The difference with SS method is that it employs another identity, in (92), for estimating the ratios $\frac{Z(\beta_k)}{Z(\beta_{k-1})}$. Figure 3 summarizes the relationships among the identities (49)-(71) and their multi-stages extensions: the SS method and path sampling scheme, respectively.

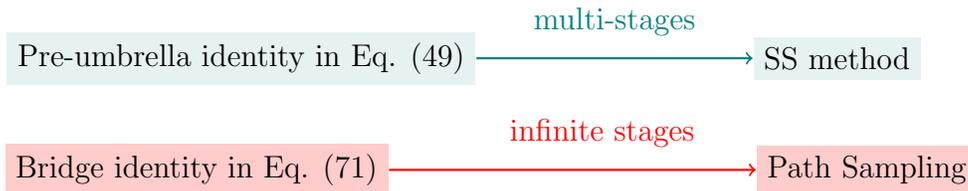


Figure 3: Relationships among the identities (49)-(71) and their multi-stages extensions: the SS method and path sampling scheme, respectively.

4.3.6 Method of Power Posteriors

The previous expression (101) can also be converted into an integral in $[0, 1]$ as follows

$$\begin{aligned}
\log Z &= \mathbb{E}_{p(\boldsymbol{\theta}, \beta | \mathbf{y})} \left[\frac{\log \ell(\mathbf{y} | \boldsymbol{\theta})}{p(\beta)} \right], \\
&= \int_0^1 d\beta \int_{\Theta} \frac{\log \ell(\mathbf{y} | \boldsymbol{\theta})}{p(\beta)} \frac{\pi(\boldsymbol{\theta} | \mathbf{y}, \beta)}{Z(\beta)} p(\beta) d\boldsymbol{\theta}, \\
&= \int_0^1 d\beta \int_{\Theta} \log \ell(\mathbf{y} | \boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta} | \mathbf{y}, \beta)}{Z(\beta)} d\boldsymbol{\theta}, \\
&= \int_0^1 \mathbb{E}_{P(\boldsymbol{\theta} | \mathbf{y}, \beta)} [\log \ell(\mathbf{y} | \boldsymbol{\theta})] d\beta,
\end{aligned} \tag{103}$$

where $P(\boldsymbol{\theta} | \mathbf{y}, \beta) = \frac{\pi(\boldsymbol{\theta} | \mathbf{y}, \beta)}{Z(\beta)}$ is a power posterior. The power posterior method aims at estimating the integral above by applying a quadrature rule. For instance, choosing a discretization $0 = \beta_0 < \beta_1 < \dots < \beta_{K-1} < \beta_K = 1$, leads to approximations of order 0,

$$\widehat{\log Z} = \sum_{k=1}^K (\beta_k - \beta_{k-1}) \mathbb{E}_{P(\boldsymbol{\theta} | \mathbf{y}, \beta_{k-1})} [\log \ell(\mathbf{y} | \boldsymbol{\theta})], \tag{104}$$

or order 1 (trapezoidal rule),

$$\widehat{\log Z} = \sum_{k=1}^K (\beta_k - \beta_{k-1}) \frac{\mathbb{E}_{P(\boldsymbol{\theta} | \mathbf{y}, \beta_k)} [\log \ell(\mathbf{y} | \boldsymbol{\theta})] + \mathbb{E}_{P(\boldsymbol{\theta} | \mathbf{y}, \beta_{k-1})} [\log \ell(\mathbf{y} | \boldsymbol{\theta})]}{2}, \tag{105}$$

where the expected values w.r.t. the power posteriors can be independently approximated via MCMC,

$$\mathbb{E}_{P(\boldsymbol{\theta} | \mathbf{y}, \beta_k)} [\log \ell(\mathbf{y} | \boldsymbol{\theta})] \approx \frac{1}{N} \sum_{i=1}^N \log \ell(\mathbf{y} | \boldsymbol{\theta}_{i,k}), \quad \{\boldsymbol{\theta}_{i,k}\}_{i=1}^N \sim P(\boldsymbol{\theta} | \mathbf{y}, \beta_k), \quad k = 0, \dots, K. \tag{106}$$

Remark 14. *The identity (103) of method of power posteriors is derived by the path sampling identity with a geometric path, as shown in (100)-(101). In this sense, the method of power posteriors is a special case of path sampling. However, unlike in path sampling, the final approximation (105) is based on a deterministic quadrature.*

Remark 15. *Note that the approximation in Eq. (105) is biased due to using a deterministic quadrature, unlike the path sampling approximation in Eq. (99) which is unbiased.*

Remark 16. *The need of using several values β_i (i.e., several tempered posteriors) seems apparent in the estimator (99)-(105). For instance, in (105), the choice a small value of K yields a poor approximation of the integral (103). This is not the case in the SS method.*

Extensions. Several improvements of the method of power posterior have been proposed in

the literature [61, 62]. In [61], the authors note that the derivative of the integrand in (103) corresponds to

$$\frac{d}{d\beta} \mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y},\beta)}[\log \ell(\mathbf{y}|\boldsymbol{\theta})] = \text{var}_{P(\boldsymbol{\theta}|\mathbf{y},\beta)}[\log \ell(\mathbf{y}|\boldsymbol{\theta})] \quad (107)$$

so they propose to use this information to refine the trapezoidal rule in (105) by adding additional terms

$$\widehat{\log Z} = \sum_{k=1}^K (\beta_k - \beta_{k-1}) \frac{\mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y},\beta_k)}[\log \ell(\mathbf{y}|\boldsymbol{\theta})] + \mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y},\beta_{k-1})}[\log \ell(\mathbf{y}|\boldsymbol{\theta})]}{2} \quad (108)$$

$$\sum_{k=1}^K \frac{(\beta_k - \beta_{k-1})^2}{12} [\text{var}_{P(\boldsymbol{\theta}|\mathbf{y},\beta_k)}[\log \ell(\mathbf{y}|\boldsymbol{\theta})] - \text{var}_{P(\boldsymbol{\theta}|\mathbf{y},\beta_{k-1})}[\log \ell(\mathbf{y}|\boldsymbol{\theta})]], \quad (109)$$

This improvement comes at no extra cost since the same MCMC samples, used to estimate the expectations in (106), can be also used to estimate the variances in (109). They also propose constructing the temperature ladder recursively, starting from $\beta_0 = 0$ and $\beta_K = 1$, by leveraging the estimates of $\mathbb{E}_{P(\boldsymbol{\theta}|\mathbf{y},\beta_k)}[\log \ell(\mathbf{y}|\boldsymbol{\theta})]$ and $\text{var}_{P(\boldsymbol{\theta}|\mathbf{y},\beta_k)}[\log \ell(\mathbf{y}|\boldsymbol{\theta})]$ (for further details see [61, Sect. 2.2]). In [62], they propose the use of control variates, a variance reduction technique, in order to improve the statistical efficiency of the estimator (105). However, this can only be applied in settings where $\nabla_{\boldsymbol{\theta}} \log P(\boldsymbol{\theta}|\mathbf{y}, \beta)$ is available.

4.3.7 On the selection of β_k

The method of power posteriors and SS sampling require setting an increasing sequence of β 's. Some strategies for selecting the sequence of values β_k 's, with $\beta_0 = 0$ and $\beta_K = 1$, are discussed, e.g., in [60, 61, 63]. A uniform sequence $\beta_k = \frac{k}{K}$ for $k = 0, \dots, K$ can be considered, although [60] recommends putting more values near $\beta = 0$, since it is where $P(\boldsymbol{\theta}|\mathbf{y}, \beta)$ is changing more rapidly. More generally, we can consider $\beta_k = (\frac{k}{K})^{1/\alpha}$. For choice of $\alpha \in [0, 1]$, the values β_k are evenly-spaced quantiles of a Beta($\alpha, 1$), concentrating more and more near $\beta = 0$ as α decreases to 0 [63].

The path sampling method requires defining a prior density $p(\beta)$ from which samples are drawn. It can be shown that, for any given path, the optimal choice of $p(\beta)$ is a generalized local Jeffreys prior [59, Sect. 4.1].

4.3.8 Connection between stepping-stone and power posteriors methods

Taking the logarithm of the SS estimator (95), we obtain

$$\log \widehat{Z}_{\text{SS}} = \sum_{k=1}^K \log \left(\frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_{i,k-1})^{\beta_k - \beta_{k-1}} \right).$$

Applying the Jensen inequality and property of the logarithm, we can write

$$\begin{aligned} \log \widehat{Z}_{\text{SS}} &\geq \sum_{k=1}^K \left(\frac{1}{N} \sum_{i=1}^N \log \ell(\mathbf{y}|\boldsymbol{\theta}_{i,k-1})^{\beta_k - \beta_{k-1}} \right), \\ &\geq \sum_{k=1}^K (\beta_k - \beta_{k-1}) \left(\frac{1}{N} \sum_{i=1}^N \log \ell(\mathbf{y}|\boldsymbol{\theta}_{i,k-1}) \right). \end{aligned}$$

The last expression is the estimator of the power posteriors method of order 0, i.e., replacing Eq. (106) into (104). If we denote here this estimator here as $\widehat{\log Z}_{PP}$, then we have $\log \widehat{Z}_{\text{SS}} \geq \widehat{\log Z}_{PP}$. Recall also the SS estimator is unbiased.

5 Advanced schemes combining MCMC and IS

In the previous sections, we have already introduced several methods which require the use of MCMC algorithms in order to draw from complex proposal densities. The RIS estimator, path sampling, power posteriors and the SS sampling schemes are some examples. All these previous schemes could be assigned to the family of “MCMC-within-IS” techniques. In this section, we describe more sophisticated schemes for estimating the evidence, which combine MCMC and IS techniques: Annealed Importance Sampling (An-IS) in Section 5.1, Sequential Monte Carlo (SMC) in Section 5.2, Multiple Try Metropolis (MTM) in Section 5.3, and Layered Adaptive importance Sampling (LAIS) in Section 5.4. An-IS and SMC can be also considered “MCMC-within-IS” techniques. They provide alternative ways to employ tempered posteriors and are related to SS method, described in the previous section. We also discuss the use of MCMC transitions and resampling steps for design efficient AIS schemes. The MTM algorithm described here is an MCMC method, which belongs to the family of “IS-within-MCMC” techniques. Indeed, internal IS steps are used for proposing good candidates as new state of the chain. LAIS is an AIS scheme driven by MCMC transitions. Since the the adaptation and sampling parts can be completely separated, LAIS can be considered as a “IS-after-MCMC” technique.

5.1 MCMC-within-IS: weighted samples after MCMC iterations

In this section, we will see how to *properly* weight samples obtained by different MCMC iterations. We denote as $K(\mathbf{z}|\boldsymbol{\theta})$ the transition kernel which summarizes all the steps of the employed MCMC algorithm. Note that generally $K(\mathbf{z}|\boldsymbol{\theta})$ cannot be evaluated. However, we can use MCMC kernels $K(\mathbf{z}|\boldsymbol{\theta})$ in the same fashion as proposal densities, considering the concept of the so-called *proper weighting* [1, 64].

5.1.1 Weighting a sample after one MCMC iteration

Let us consider the following procedure:

1. Draw $\boldsymbol{\theta}_0 \sim q(\boldsymbol{\theta})$ (where $q(\boldsymbol{\theta})$ is normalized, for simplicity).

2. Draw $\boldsymbol{\theta}_1 \sim K(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)$, where the kernel K leaves invariant density $\bar{\eta}(\boldsymbol{\theta}) = \frac{1}{c}\eta(\boldsymbol{\theta})$, i.e.,

$$\int_{\Theta} K(\boldsymbol{\theta}'|\boldsymbol{\theta})\bar{\eta}(\boldsymbol{\theta})d\boldsymbol{\theta} = \bar{\eta}(\boldsymbol{\theta}'). \quad (110)$$

3. Assign to $\boldsymbol{\theta}_1$ the weight

$$\rho(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{\eta(\boldsymbol{\theta}_0) \pi(\boldsymbol{\theta}_1|\mathbf{y})}{q(\boldsymbol{\theta}_0) \eta(\boldsymbol{\theta}_1)}. \quad (111)$$

This weight is *proper* in the sense that can be used for building unbiased estimator Z (or other moments $P(\boldsymbol{\theta}|\mathbf{y})$), as described in the Liu's definition [2, Section 14.2], [1, Section 2.5.4]. Indeed, we can write

$$\begin{aligned} \mathbb{E}[\rho(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)] &= \int_{\Theta} \int_{\Theta} \rho(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) K(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0) q(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_1, \\ &= \int_{\Theta} \int_{\Theta} \frac{\eta(\boldsymbol{\theta}_0) \pi(\boldsymbol{\theta}_1)}{q(\boldsymbol{\theta}_0) \eta(\boldsymbol{\theta}_1)} K(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0) q(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_1, \\ &= \int_{\Theta} \frac{\pi(\boldsymbol{\theta}_1)}{\eta(\boldsymbol{\theta}_1)} \left[\int_{\Theta} \eta(\boldsymbol{\theta}_0) K(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 \right] d\boldsymbol{\theta}_1, \\ &= \int_{\Theta} \frac{\pi(\boldsymbol{\theta}_1)}{c\bar{\eta}(\boldsymbol{\theta}_1)} c\bar{\eta}(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 = \int_{\Theta} \pi(\boldsymbol{\theta}_1|\mathbf{y}) d\boldsymbol{\theta}_1 = Z. \end{aligned} \quad (112)$$

Note that if $\eta(\boldsymbol{\theta}) \equiv \pi(\boldsymbol{\theta}|\mathbf{y})$ then $\rho(\boldsymbol{\theta}_1) = \frac{\pi(\boldsymbol{\theta}_0|\mathbf{y})}{q(\boldsymbol{\theta}_0)}$, i.e., the IS weights remain unchanged after an MCMC iteration with invariant density $\pi(\boldsymbol{\theta}|\mathbf{y})$. Hence, if we repeat the procedure above N times generating $\{\boldsymbol{\theta}_0^{(n)}, \boldsymbol{\theta}_1^{(n)}\}_{n=1}^N$, we can build the following unbiased estimator of the Z ,

$$\hat{Z} = \frac{1}{N} \sum_{n=1}^N \rho(\boldsymbol{\theta}_0^{(n)}, \boldsymbol{\theta}_1^{(n)}) = \frac{1}{N} \sum_{n=1}^N \frac{\eta(\boldsymbol{\theta}_0^{(n)}) \pi(\boldsymbol{\theta}_1^{(n)}|\mathbf{y})}{q(\boldsymbol{\theta}_0^{(n)}) \eta(\boldsymbol{\theta}_1^{(n)})} \quad (113)$$

In the next section, we extend this idea where different MCMC updates are applied, each one addressing a different invariant density.

5.1.2 Annealed Importance Sampling (An-IS)

In the previous section, we have considered the application of one MCMC kernel $K(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)$ (that could be formed by different MCMC steps). Below, we consider the application of several MCMC kernels addressing different target pdfs, and show their consequence in the weighting strategy. We consider again a sequence of tempered versions of the posterior, $\pi_1(\boldsymbol{\theta}|\mathbf{y}), \pi_2(\boldsymbol{\theta}|\mathbf{y}), \dots, \pi_L(\boldsymbol{\theta}|\mathbf{y}) \equiv \pi(\boldsymbol{\theta}|\mathbf{y})$, where the L -th version, $\pi_L(\boldsymbol{\theta}|\mathbf{y})$, coincides with the target function $\pi(\boldsymbol{\theta}|\mathbf{y})$. One possibility is to consider $\pi_i(\boldsymbol{\theta}|\mathbf{y}) = [\pi(\boldsymbol{\theta}|\mathbf{y})]^{\beta_i} = g(\boldsymbol{\theta})^{\beta_i} \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta_i}$ or tempered posteriors,

$$\pi_i(\boldsymbol{\theta}|\mathbf{y}) = g(\boldsymbol{\theta}) \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta_i} \quad \text{where} \quad 0 \leq \beta_1 \leq \beta_2 \leq \dots \leq \beta_L = 1. \quad (114)$$

as in path sampling and power posteriors. In any case, smaller β values correspond to flatter distributions.⁴ The use of the tempered sequence of target pdfs usually improve the mixing of the algorithm and foster the exploration of the space Θ . Since only the last function is the true target, $\pi_L(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}|\mathbf{y})$, different schemes have been proposed for suitable weighting the final samples.

Let us consider conditional $L - 1$ kernels $K_i(\mathbf{z}|\boldsymbol{\theta})$ (with $L \geq 2$), representing the probability of different MCMC updates of jumping from the state $\boldsymbol{\theta}$ to the state \mathbf{z} (note that each K_i can summarize the application of several MCMC steps), each one leaving invariant a different tempered target, $P_i(\boldsymbol{\theta}|\mathbf{y}) \propto \pi_i(\boldsymbol{\theta}|\mathbf{y})$. The Annealed Importance Sampling (An-IS) is given in Table 10. Note

Table 10: Annealed Importance Sampling (An-IS)

1. Draw N samples $\boldsymbol{\theta}_0^{(n)} \sim P_0(\boldsymbol{\theta}|\mathbf{y})$ (usually $g(\boldsymbol{\theta})$) for $n = 1, \dots, N$.
2. For $k = 1, \dots, L - 1$:
 - (a) Draw $\boldsymbol{\theta}_k^{(n)} \sim K_k(\boldsymbol{\theta}|\boldsymbol{\theta}_{k-1}^{(n)})$ leaving invariant $P_k(\boldsymbol{\theta}|\mathbf{y})$ for $n = 1, \dots, N$, i.e., we generate N samples using an MCMC with invariant distribution $P_k(\boldsymbol{\theta}|\mathbf{y})$ (with different starting points $\boldsymbol{\theta}_{k-1}^{(n)}$).
 - (b) Compute the weight associated to the sample $\boldsymbol{\theta}_k^{(n)}$, for $n = 1, \dots, N$,

$$\rho_k^{(n)} = \prod_{i=0}^k \frac{\pi_{i+1}(\boldsymbol{\theta}_i^{(n)}|\mathbf{y})}{\pi_i(\boldsymbol{\theta}_i^{(n)}|\mathbf{y})} = \rho_{k-1}^{(n)} \frac{\pi_{k+1}(\boldsymbol{\theta}_k^{(n)}|\mathbf{y})}{\pi_k(\boldsymbol{\theta}_k^{(n)}|\mathbf{y})}. \quad (115)$$

3. Return the weighted sample $\{\boldsymbol{\theta}_{L-1}^{(n)}, \rho_{L-1}^{(n)}\}_{n=1}^N$. The estimator of the marginal likelihood is

$$\widehat{Z} = \frac{1}{N} \sum_{n=1}^N \rho_{L-1}^{(n)}.$$

Combinations of An-IS with path sampling and power posterior methods can be also considered, employing the information of the rest of intermediate densities.

that, when $L = 2$, we have $\rho_1^{(n)} = \frac{\pi_1(\boldsymbol{\theta}_0^{(n)}|\mathbf{y})}{q(\boldsymbol{\theta}_0^{(n)})} \frac{\pi(\boldsymbol{\theta}_1^{(n)}|\mathbf{y})}{\pi_1(\boldsymbol{\theta}_1^{(n)}|\mathbf{y})}$. If, $\pi_1 = \pi_2 = \dots = \pi_{L-1} = \eta \neq \pi$, then the weight is $\rho_{L-1} = \frac{\eta(\boldsymbol{\theta}_0^{(n)})}{P_0(\boldsymbol{\theta}_0^{(n)}|\mathbf{y})} \frac{\pi(\boldsymbol{\theta}_{L-1}^{(n)}|\mathbf{y})}{\eta(\boldsymbol{\theta}_{L-1}^{(n)})}$.

The method above can be modified by incorporating an additional MCMC transition $\boldsymbol{\theta}_L \sim K_L(\boldsymbol{\theta}|\boldsymbol{\theta}_{L-1})$, which leaves invariant $P_L(\boldsymbol{\theta}|\mathbf{y}) = P(\boldsymbol{\theta}|\mathbf{y})$. However, since $P_L(\boldsymbol{\theta}|\mathbf{y})$ is the true target pdf, as we have seen above the weight remains unchanged (see the case $\bar{\eta}(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$ in the previous section). Hence, in this scenario, the output would be $\{\boldsymbol{\theta}_L^{(n)}, \rho_L^{(n)}\} = \{\boldsymbol{\theta}_L^{(n)}, \rho_{L-1}^{(n)}\}$, i.e.,

⁴Another alternative is to use the so-called *data tempering* [65], for instance, setting $\pi_i(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}|y_1, \dots, y_{d+i})$, where $d \geq 1$ and $d + L = D_y$ (recall that $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$).

$\rho_L^{(n)} = \rho_{L-1}^{(n)}$. This method has been proposed in [66] but similarly schemes can be found in [65, 67].

Remark 17. *The stepping-stones (SS) sampling method described in Section 4.3 is strictly connected to an Ann-IS scheme. See Figures 2 and 4 for a comparison of the sampling procedures.*

Interpretation as Standard IS. For the sake of simplicity, here we consider *reversible* kernels, i.e., each kernel satisfies the detailed balance condition

$$\pi_i(\boldsymbol{\theta}|\mathbf{y})K_i(\mathbf{z}|\boldsymbol{\theta}) = \pi_i(\mathbf{z}|\mathbf{y})K_i(\boldsymbol{\theta}|\mathbf{z}) \quad \text{so that} \quad \frac{K_i(\mathbf{z}|\boldsymbol{\theta})}{K_i(\boldsymbol{\theta}|\mathbf{z})} = \frac{\pi_i(\mathbf{z}|\mathbf{y})}{\pi_i(\boldsymbol{\theta}|\mathbf{y})}. \quad (116)$$

We show that the weighting strategy suggested by An-IS can be interpreted as a standard IS weighting considering the following extended target density, defined in the extended space Θ^L ,

$$\pi_g(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{L-1}|\mathbf{y}) = \pi(\boldsymbol{\theta}_{L-1}|\mathbf{y}) \prod_{k=1}^{L-1} K_k(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_k). \quad (117)$$

Note that π_g has the true target π as a marginal pdf. Let also consider an extended proposal pdf defined as

$$q_g(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{L-1}) = P_0(\boldsymbol{\theta}_0|\mathbf{y}) \prod_{k=1}^{L-1} K_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}). \quad (118)$$

The standard IS weight of an extended sample $[\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{L-1}]$ in the extended space Θ^L is

$$w(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{L-1}) = \frac{\pi_g(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{L-1}|\mathbf{y})}{q_g(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{L-1})} = \frac{\pi(\boldsymbol{\theta}_{L-1}|\mathbf{y}) \prod_{k=1}^{L-1} K_k(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_k)}{P_0(\boldsymbol{\theta}_0|\mathbf{y}) \prod_{k=1}^{L-1} K_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})}. \quad (119)$$

Replacing the expression $\frac{K_i(\mathbf{z}|\boldsymbol{\theta})}{K_i(\boldsymbol{\theta}|\mathbf{z})} = \frac{\pi_i(\mathbf{z}|\mathbf{y})}{\pi_i(\boldsymbol{\theta}|\mathbf{y})}$ in (119), we obtain the Ann-IS weights

$$w(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{L-1}) = \frac{\pi(\boldsymbol{\theta}_{L-1}|\mathbf{y})}{P_0(\boldsymbol{\theta}_0|\mathbf{y})} \prod_{k=1}^{L-1} \frac{\pi_k(\boldsymbol{\theta}_{k-1}|\mathbf{y})}{\pi_k(\boldsymbol{\theta}_k|\mathbf{y})}, \quad (120)$$

$$= \frac{\pi_1(\boldsymbol{\theta}_0|\mathbf{y})}{P_0(\boldsymbol{\theta}_0|\mathbf{y})} \prod_{k=1}^{L-1} \frac{\pi_{k+1}(\boldsymbol{\theta}_k|\mathbf{y})}{\pi_k(\boldsymbol{\theta}_k|\mathbf{y})} = \prod_{k=0}^{L-1} \frac{\pi_{k+1}(\boldsymbol{\theta}_k|\mathbf{y})}{\pi_k(\boldsymbol{\theta}_k|\mathbf{y})} = \rho_{L-1}, \quad (121)$$

where we have used $\pi_L(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}|\mathbf{y})$ and just rearranged the numerator. The sampling procedure in An-IS is graphically represented in Figure 4.

5.2 Weighted samples after MCMC and resampling steps

In this section, we consider also the use of resampling steps jointly with MCMC transitions. The resulting algorithm is quite sophisticated (formed by several components that should be chosen by the user) but it is a very general technique, which includes the classical particle filters, several adaptive IS (AIS) schemes and the An-IS method as special case [68].

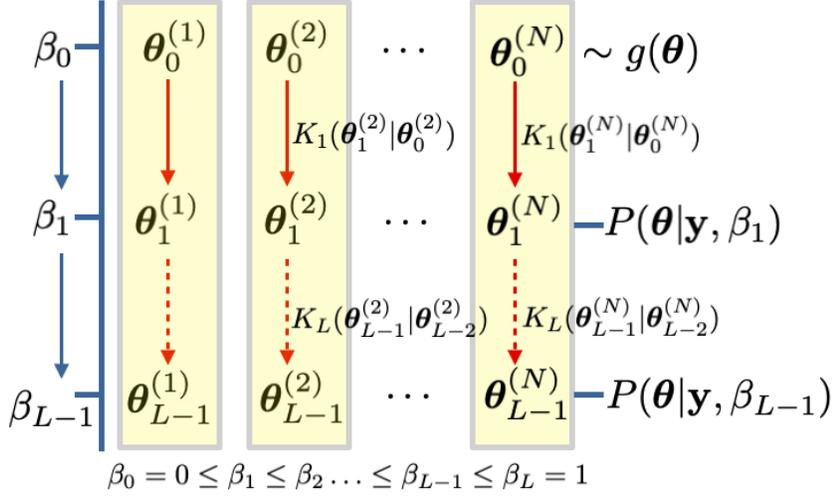


Figure 4: Sampling procedure in the An-IS method.

5.2.1 Generic Sequential Monte Carlo

In this section, we describe a sequential IS scheme which encompasses the previous Ann-IS algorithm as a special case. The method described here uses jointly MCMC transitions and, additionally, resampling steps as well. It is called Sequential Monte Carlo (SMC), since we have a sequence of target pdfs $\pi_k(\boldsymbol{\theta}|\mathbf{y})$, $k = 1, \dots, L$ [68]. This sequence of target densities can be defined by a state-space model as in a classical particle filtering framework (truly sequential scenario, where the goal is to track dynamic parameters). Alternatively, we can also consider a static scenario as in the previous sections, i.e., the resulting algorithm is an iterative importance sampler where we consider a sequence of *tempered* densities $\pi_k(\boldsymbol{\theta}|\mathbf{y}) = g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^{\beta_k}$, where $0 \leq \beta_1 \leq \dots \leq \beta_L = 1$, as in Eq.(114), so that $\pi_L(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}|\mathbf{y})$ [68]. Let us again define an extended proposal density in the domain Θ^k ,

$$\tilde{q}_k(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) = q_1(\boldsymbol{\theta}_1) \prod_{i=2}^k F_i(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1}) : \Theta^k \rightarrow \mathbb{R}, \quad (122)$$

where $q_1(\boldsymbol{\theta}_1)$ is a marginal proposal and $F_i(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})$ are generic forward transition pdfs, that will be used as partial proposal pdfs. Extending the space from Θ^k to Θ^{k+1} (increasing its dimension), note that we can write the recursive equation

$$\tilde{q}_{k+1}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1}) = F_{k+1}(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k)\tilde{q}_k(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) : \Theta^{k+1} \rightarrow \mathbb{R}.$$

The marginal proposal pdfs are

$$\begin{aligned}
q_k(\boldsymbol{\theta}_k) &= \int_{\Theta^{k-1}} \tilde{q}_k(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) d\boldsymbol{\theta}_{1:k-1} \\
&= \int_{\Theta^{k-1}} q_1(\boldsymbol{\theta}_1) \prod_{i=2}^k F_i(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{i-1}) d\boldsymbol{\theta}_{1:k-1}, \tag{123}
\end{aligned}$$

$$\begin{aligned}
&= \int_{\Theta} \left[\int_{\Theta^{k-2}} q_1(\boldsymbol{\theta}_1) \prod_{i=2}^k F_i(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{i-1}) d\boldsymbol{\theta}_{1:k-2} \right] F_k(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) d\boldsymbol{\theta}_{k-1}, \\
&= \int_{\Theta} q_{k-1}(\boldsymbol{\theta}_{k-1}) F_k(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) d\boldsymbol{\theta}_{k-1}, \tag{124}
\end{aligned}$$

Therefore, we would be interested in computing the *marginal* IS weights, $w_k = \frac{\pi_k(\boldsymbol{\theta}_k | \mathbf{y})}{q_k(\boldsymbol{\theta}_k)}$, for each k . However note that, in general, the marginal proposal pdfs $q_k(\boldsymbol{\theta}_k)$ cannot be computed and then cannot be evaluated. A suitable alternative approach is described next. Let us consider the extended target pdf defined as

$$\tilde{\pi}_k(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k | \mathbf{y}) = \pi_k(\boldsymbol{\theta}_k | \mathbf{y}) \prod_{i=2}^k B_{i-1}(\boldsymbol{\theta}_{i-1} | \boldsymbol{\theta}_i) : \Theta^k \rightarrow \mathbb{R}, \tag{125}$$

$B_{i-1}(\boldsymbol{\theta}_{i-1} | \boldsymbol{\theta}_i)$ are arbitrary backward transition pdfs. Note that the space of $\{\tilde{\pi}_k\}$ increases as k grows, and π_k is always a marginal pdf of $\tilde{\pi}_k$. Moreover, writing the previous equation for $k+1$

$$\tilde{\pi}_{k+1}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1} | \mathbf{y}) = \pi_{k+1}(\boldsymbol{\theta}_{k+1} | \mathbf{y}) \prod_{i=2}^{k+1} B_{i-1}(\boldsymbol{\theta}_{i-1} | \boldsymbol{\theta}_i),$$

and writing the ratio of both, we get

$$\frac{\tilde{\pi}_{k+1}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1} | \mathbf{y})}{\tilde{\pi}_k(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k | \mathbf{y})} = \frac{\pi_{k+1}(\boldsymbol{\theta}_{k+1} | \mathbf{y})}{\pi_k(\boldsymbol{\theta}_k | \mathbf{y})} B_k(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k+1}). \tag{126}$$

Therefore, the IS weights in the extended space Θ^k are

$$w_k = \frac{\tilde{\pi}_k(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k | \mathbf{y})}{\tilde{q}_k(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k)} \tag{127}$$

$$= \frac{\tilde{\pi}_{k-1}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{k-1} | \mathbf{y}) \frac{\pi_k(\boldsymbol{\theta}_k | \mathbf{y})}{\pi_{k-1}(\boldsymbol{\theta}_{k-1} | \mathbf{y})} B_{k-1}(\boldsymbol{\theta}_{k-1} | \boldsymbol{\theta}_k)}{\tilde{q}_{k-1}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{k-1}) F_k(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1})}, \tag{128}$$

$$= w_{k-1} \frac{\pi_k(\boldsymbol{\theta}_k | \mathbf{y}) B_{k-1}(\boldsymbol{\theta}_{k-1} | \boldsymbol{\theta}_k)}{\pi_{k-1}(\boldsymbol{\theta}_{k-1} | \mathbf{y}) F_k(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1})}. \tag{129}$$

where we have replaced $w_{k-1} = \frac{\tilde{\pi}_{k-1}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{k-1} | \mathbf{y})}{\tilde{q}_{k-1}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{k-1})}$. The recursive formula in Eq. (129) is the key expression for several sequential IS techniques. The SMC scheme summarized in Table 11 is a general framework which contains different algorithms as a special cases [68]. In Table 11, we have used the notation $\boldsymbol{\theta}_{1:k} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k]$.

Table 11: Generic Sequential Monte Carlo (SMC)

1. Draw $\boldsymbol{\theta}_1^{(n)} \sim q_1(\boldsymbol{\theta})$, $n = 1, \dots, N$.

2. For $k = 2, \dots, L$:

(a) Draw N samples $\boldsymbol{\theta}_k^{(n)} \sim F_k(\boldsymbol{\theta}|\boldsymbol{\theta}_{k-1}^{(n)})$.

(b) Compute the weights

$$w_k^{(n)} = w_{k-1}^{(n)} \frac{\pi_k(\boldsymbol{\theta}_k^{(n)}|\mathbf{y})B_{k-1}(\boldsymbol{\theta}_{k-1}^{(n)}|\boldsymbol{\theta}_k^{(n)})}{\pi_{k-1}(\boldsymbol{\theta}_{k-1}^{(n)}|\mathbf{y})F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(n)})}, \quad (130)$$

$$= w_{k-1}^{(n)} \gamma_k^{(n)}, \quad , k = 1, \dots, L, \quad (131)$$

where we set $\gamma_k^{(n)} = \frac{\pi_k(\boldsymbol{\theta}_k^{(n)}|\mathbf{y})B_{k-1}(\boldsymbol{\theta}_{k-1}^{(n)}|\boldsymbol{\theta}_k^{(n)})}{\pi_{k-1}(\boldsymbol{\theta}_{k-1}^{(n)}|\mathbf{y})F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(n)})}$.

(c) Normalize the weights $\bar{w}_k^{(n)} = \frac{w_k^{(n)}}{\sum_{j=1}^N w_k^{(j)}}$, for $n = 1, \dots, N$.

(d) If $\widehat{ESS} \leq \epsilon N$:

(with $0 \leq \epsilon \leq 1$ and \widehat{ESS} is a effective sample size measure [69], see section 5.2.2)

- i. Resample N times $\{\boldsymbol{\theta}_{1:k}^{(1)}, \dots, \boldsymbol{\theta}_{1:k}^{(N)}\}$ according to $\{\bar{w}_k^{(n)}\}_{n=1}^N$, obtaining $\{\bar{\boldsymbol{\theta}}_{1:k}^{(1)}, \dots, \bar{\boldsymbol{\theta}}_{1:k}^{(N)}\}$.
- ii. Set $\boldsymbol{\theta}_{1:k}^{(n)} = \bar{\boldsymbol{\theta}}_{1:k}^{(n)}$, $\widehat{Z}_k = \frac{1}{N} \sum_{n=1}^N w_k^{(n)}$ and $w_k^{(n)} = \widehat{Z}_k$ for all $n = 1, \dots, N$ [70, 64, 71, 20].

3. Return the cloud of weighted particles and

$$\widehat{Z} = \widehat{Z}_L = \frac{1}{N} \sum_{n=1}^N w_L^{(n)},$$

if a proper weighting of the resampled particles is used (as suggested in the step 2(d)-ii above). Otherwise, you can use another estimator \widehat{Z}_L , as shown in Section 5.2.2 and the Supplementary Material.

Choice of the forward functions. One possible choice is to use independent proposal pdfs, i.e., $F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) = F_k(\boldsymbol{\theta}_k)$ or random walk proposal $F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})$, where F_k represents standard distributions (e.g., Gaussian or t-Student). An alternative is to choose $F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) = K_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})$, i.e., an MCMC kernel with invariant pdf $P_k(\boldsymbol{\theta}_k|\mathbf{y})$.

Choice of backward functions. It is possible to show that the optimal backward transitions $\{B_k\}_{k=1}^L$ are [68]

$$B_{k-1}(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_k) = \frac{q_{k-1}(\boldsymbol{\theta}_{k-1})}{q_k(\boldsymbol{\theta}_k)} F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}). \quad (132)$$

This choice reduces the variance of the weights [68]. However, generally, the marginal proposal q_k in Eq. (123) cannot be computed (are not available), other possible $\{B_k\}$ should be considered. For instance, with the choice

$$B_{k-1}(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_k) = \frac{\pi_k(\boldsymbol{\theta}_{k-1}|\mathbf{y})}{\pi_k(\boldsymbol{\theta}_k|\mathbf{y})} F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}), \quad (133)$$

we obtain

$$w_k = w_{k-1} \frac{\pi_k(\boldsymbol{\theta}_k|\mathbf{y}) \frac{\pi_k(\boldsymbol{\theta}_{k-1}|\mathbf{y})}{\pi_k(\boldsymbol{\theta}_k|\mathbf{y})} F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})}{\pi_{k-1}(\boldsymbol{\theta}_{k-1}|\mathbf{y}) F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})} \quad (134)$$

$$= w_{k-1} \frac{\pi_k(\boldsymbol{\theta}_{k-1}|\mathbf{y})}{\pi_{k-1}(\boldsymbol{\theta}_{k-1}|\mathbf{y})}, \quad (135)$$

which is exactly the update rule for the weights in An-IS.

Remark 18. With the choice of $B_{k-1}(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_k)$ as in Eq. 133, and if $F_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) = K_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})$ is an MCMC kernel with invariant $P_k(\boldsymbol{\theta}_k|\mathbf{y})$, then we come back to An-IS algorithm [66, 65, 67], described in Table 10. Hence, the An-IS scheme is a special case of SMC method.

Several other methods are contained as special cases of algorithm in Table 11, with specific choice of $\{B_k\}$, $\{K_k\}$ and $\{\pi_k\}$, e.g., the Population Monte Carlo (PMC) method [72], that is a well-known AIS scheme. The sampling procedure in SMC is graphically represented in Figure 5.

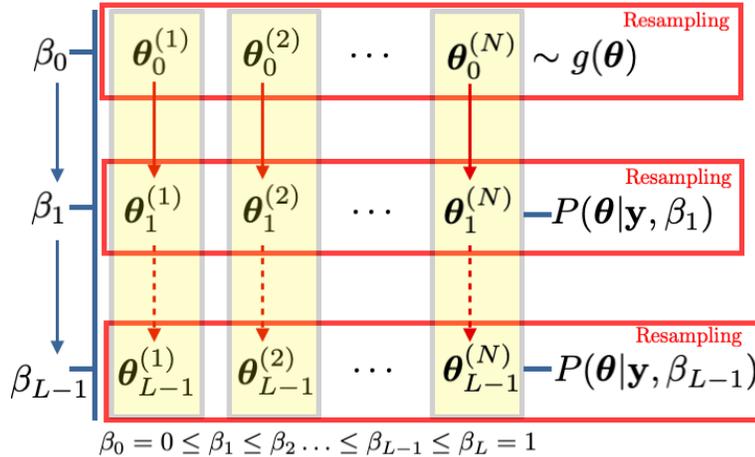


Figure 5: Sampling procedure in SMC. In this figure, we have considered resampling steps at each iteration ($\epsilon = 1$).

5.2.2 Evidence computation in a sequential framework with resampling steps

The generic algorithm in Table 11 employs also resampling steps. Resampling consists in drawing particles from the current cloud according to the normalized importance weights $\bar{w}_k^{(n)}$, for $n = 1, \dots, N$. The resampling steps are applied only in certain iterations taking into account an ESS approximation, such as $\widehat{ESS} = \frac{1}{\sum_{n=1}^N (\bar{w}_k^{(n)})^2}$, or $\widehat{ESS} = \frac{1}{\max_n \bar{w}_k^{(n)}}$ [73, 69]. Generally, if $\frac{1}{N} \widehat{ESS}$ is smaller than a pre-established threshold $\epsilon \in [0, 1]$, all the particles are resampled. Thus, the condition for the adaptive resampling can be expressed as $\widehat{ESS} < \epsilon N$. When $\epsilon = 1$, the resampling is applied at each iteration [74, 75]. If $\epsilon = 0$, no resampling steps are applied, and we have a simple sequential importance sampling (SIS) method. There are two possible estimators of Z_k in a sequential scenario:

$$\widehat{Z}_k^{(1)} = \frac{1}{N} \sum_{n=1}^N w_k^{(n)} = \frac{1}{N} \sum_{n=1}^N w_{k-1}^{(n)} \gamma_k^{(n)} = \frac{1}{N} \sum_{n=1}^N \left[\prod_{j=1}^k \gamma_j^{(n)} \right], \quad (136)$$

and

$$\widehat{Z}_k^{(2)} = \prod_{j=1}^k \left[\sum_{n=1}^N \bar{w}_{j-1}^{(n)} \gamma_j^{(n)} \right]. \quad (137)$$

These two estimators are equivalent in SIS ($\epsilon = 0$, i.e., SMC without resampling), i.e., they are the same estimator, $\widehat{Z}_k^{(1)} = \widehat{Z}_k^{(2)}$. In SMC with $\epsilon > 0$ and a proper weighting of the resampled particles, as used in Table 11, the two estimators are equivalent as well [70, 64, 20]. If the proper weighting of the resampled particles is not employed, $\widehat{Z}_k^{(2)}$ is the only valid option. See Table 12 for a summary and the Supp. Material for more details.

Table 12: Possible estimators of the evidence in a sequential scenario.

Scenario	Resampling	Proper Weighting [70]	$\widehat{Z}_k^{(1)}$	$\widehat{Z}_k^{(2)}$	Equivalence
SMC - $\epsilon = 0$ (SIS)	x	—	✓	✓	✓
SMC - $\epsilon > 0$	✓	x	x	✓	x
SMC - $\epsilon > 0$	✓	✓	✓	✓	✓

5.3 IS-within-MCMC: Estimation based on Multiple Try MCMC schemes

The Multiple Try Metropolis (MTM) methods are advanced MCMC algorithms which consider different candidates as possible new state of the chain [35, 76, 77]. More specifically, at each iteration different samples are generated and compared by using some proper weights. Then one of them is selected and tested as possible future state. The main advantage of these algorithms is that they foster the exploration of a larger portion of the sample space, decreasing the correlation among the states of the generated chain. Here, we consider the use of importance weights for comparing

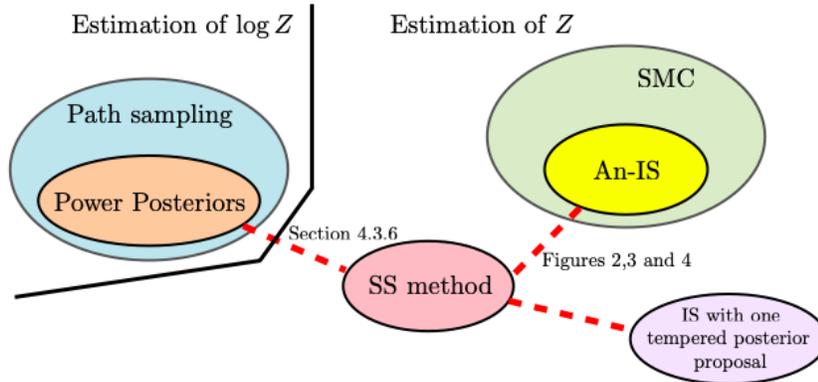


Figure 6: Graphical summary of the methods using tempered posteriors.

the different candidates, in order to provide also an estimation of the marginal likelihood [76]. More specifically, we consider the Independent Multiple Try Metropolis type 2 (IMTM-2) scheme [35] with an adaptive proposal pdf. The algorithm is given in Table 13. The mean vector and covariance matrix are adapted using the empirical estimators yielded by all the weighted candidates drawn so far, i.e., $\{\mathbf{z}_{n,\tau}, w_{n,\tau}\}$ for all $n = 1, \dots, N$ and $\tau = 1, \dots, T$. Two possible estimators of the marginal likelihood can be constructed, one based on a standard adaptive importance sampling argument $\hat{Z}^{(2)}$ [78, 79] and other based on a group importance sampling idea provided in [64].

For the sake of simplicity, we have described an independent MTM scheme, with the additional adaptation of the proposal. Random walk proposal pdfs can be also employed in an MTM algorithm [35]. In that case, the adaptation of the proposal could be not needed. However, in this scenario, the MTM algorithm requires the sampling (and weighting) of $N - 1$ additional auxiliary points. Hence, the total number of weighted samples at each iterations are $2N - 1$. These additional samples are just required for ensuring the ergodicity of the chain (including them in the acceptance probability α), but are not included as states of the Markov chain. But, for our purpose, they can be employed in the estimators of Z , as we suggest for the N candidates, $\{\mathbf{z}_{n,t}, w_{n,t}\}$, in Table 13. Note that the use of a random walk proposal in an MTM scheme of type in Table 13, could be considered as “MCMC-driven IS” method, similar to the method introduced in the next section.

5.4 IS-after-MCMC: Layered Adaptive Importance Sampling (LAIS)

The LAIS algorithm consider the use of N parallel (independent or interacting) MCMC chains with invariant pdf $P(\boldsymbol{\theta}|\mathbf{y})$ or a tempered version $P(\boldsymbol{\theta}|\beta)$ [80, 78]. Each MCMC chain can address a different tempered version $P(\boldsymbol{\theta}|\mathbf{y}, \beta)$ (or simply the posterior $P(\boldsymbol{\theta}|\mathbf{y})$) without jeopardizing the consistency of final estimators. After T iterations of the N MCMC schemes (upper layer), the resulting NT samples, $\{\boldsymbol{\mu}_{n,t}\}$, for $n = 1, \dots, N$ and $t = 1, \dots, T$ are used as location parameters of NT proposal densities $q(\boldsymbol{\theta}|\boldsymbol{\mu}_{n,t}, \mathbf{C})$. Then, these proposal pdfs are employed within a MIS scheme (lower layer), weighting the generated samples $\boldsymbol{\theta}_{n,t}$'s with the generic weight $w_{n,t} = \frac{\pi(\boldsymbol{\theta}_{n,t}|\mathbf{y})}{\Phi(\boldsymbol{\theta}_{n,t})}$ [51, 57]. In the numerator of these weights in the lower layer, we have always the unnormalized posterior

Table 13: Adaptive Independent Multiple Try Metropolis type 2 (AIMTM-2)

1. Choose the initial parameters $\boldsymbol{\mu}_t$, \mathbf{C}_t of the proposal q , an initial state $\boldsymbol{\theta}_0$ and a first estimation of the marginal likelihood \widehat{Z}_0 .

2. For $t = 1, \dots, T$:

(a) Draw $\mathbf{z}_{1,t}, \dots, \mathbf{z}_{N,t} \sim q(\mathbf{z}|\boldsymbol{\mu}_t, \mathbf{C}_t)$.

(b) Compute the importance weights $w_{n,t} = \frac{\pi(\mathbf{z}_{n,t}|\mathbf{y})}{q(\mathbf{z}_{n,t}|\boldsymbol{\mu}_t, \mathbf{C}_t)}$, for $n = 1, \dots, N$.

(c) Normalize them $\bar{w}_{n,t} = \frac{w_{n,t}}{N\widehat{Z}'}$ where

$$\widehat{Z}' = \frac{1}{N} \sum_{i=1}^N w_{i,t}, \quad \text{and set} \quad R_t = \widehat{Z}'. \quad (138)$$

(d) Resample $\boldsymbol{\theta}' \in \{\mathbf{z}_{1,t}, \dots, \mathbf{z}_{N,t}\}$ according to \bar{w}_n , with $n = 1, \dots, N$.

(e) Set $\boldsymbol{\theta}_t = \boldsymbol{\theta}'$ and $\widehat{Z}_t = \widehat{Z}'$ with probability

$$\alpha = \min \left[1, \frac{\widehat{Z}'}{\widehat{Z}_{t-1}} \right] \quad (139)$$

otherwise set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ and $\widehat{Z}_t = \widehat{Z}_{t-1}$.

(f) Update $\boldsymbol{\mu}_t$, \mathbf{C}_t computing the corresponding empirical estimators using $\{\mathbf{z}_{n,\tau}, w_{n,\tau}\}$ for all $n = 1, \dots, N$ and $\tau = 1, \dots, T$.

3. Return the chain $\{\boldsymbol{\theta}_t\}_{t=1}^T$, $\{\widehat{Z}_t\}_{t=1}^T$ and $\{R_t\}_{t=1}^T$. Two possible estimators of Z can be constructed:

$$\widehat{Z}^{(1)} = \frac{1}{T} \sum_{t=1}^T \widehat{Z}_t, \quad \widehat{Z}^{(2)} = \frac{1}{T} \sum_{t=1}^T R_t. \quad (140)$$

$\pi(\boldsymbol{\theta}_{n,t}|\mathbf{y})$. The denominator $\Phi(\boldsymbol{\theta}_{n,t})$ is a mixture of (all or a subset of) proposal densities which specifies the type of MIS scheme applied [51, 57]. The algorithm, with different possible choices of $\Phi(\boldsymbol{\theta}_{n,t})$, is shown in Table 14. The first choice in (142) is the most costly since we have to evaluate all the proposal pdfs in all the generated samples $\boldsymbol{\theta}_{n,t}$'s, but provides the best performance in terms of efficiency of the final estimator. The second and third choices are temporal and spatial mixtures, respectively. The last choice corresponds to standard importance weights given in Section 4.

Let assume $P_n(\boldsymbol{\theta}|\mathbf{y}) = P(\boldsymbol{\theta}|\mathbf{y})$ for all n in the upper layer. Considering also standard parallel Metropolis-Hastings chains in the upper layer, the number of posterior evaluations in LAIS is $2NT$. Thus, if only one chain $N = 1$ is employed in the upper layer, the number of posterior evaluations is $2T$.

Special case with recycling samples. The method in [81] can be considered as a special case

Table 14: Layered Adaptive Importance Sampling (LAIS)

1. Generate NT samples, $\{\boldsymbol{\mu}_{n,t}\}$, using N parallel MCMC chains of length T , each MCMC method using a proposal pdf $\varphi_n(\boldsymbol{\mu}|\boldsymbol{\mu}_{t-1})$, with invariant distributions a power posterior $P_n(\boldsymbol{\theta}|\mathbf{y}) = P(\boldsymbol{\theta}|\mathbf{y}, \beta_n)$ (with $\beta_n > 0$) or a posterior pdf with a smaller number of data.
2. Draw NT samples $\boldsymbol{\theta}_{n,t} \sim q(\boldsymbol{\theta}|\boldsymbol{\mu}_{n,t}, \mathbf{C})$ where $\boldsymbol{\mu}_{n,t}$ plays the role of the mean, and \mathbf{C} is a covariance matrix.
3. Assign to $\boldsymbol{\theta}_{n,t}$ the weights

$$w_{n,t} = \frac{\pi(\boldsymbol{\theta}_{n,t}|\mathbf{y})}{\Phi(\boldsymbol{\theta}_{n,t})}. \quad (141)$$

There are different possible choices for $\Phi(\boldsymbol{\theta}_{n,t})$, for instance:

$$\Phi(\boldsymbol{\theta}_{n,t}) = \frac{1}{NT} \sum_{k=1}^T \sum_{i=1}^N q_{i,k}(\boldsymbol{\theta}_{n,t}|\boldsymbol{\mu}_{i,k}, \mathbf{C}), \quad (142)$$

$$\Phi(\boldsymbol{\theta}_{n,t}) = \frac{1}{T} \sum_{k=1}^T q(\boldsymbol{\theta}_{n,t}|\boldsymbol{\mu}_{n,k}, \mathbf{C}), \quad (143)$$

$$\Phi(\boldsymbol{\theta}_{n,t}) = \frac{1}{N} \sum_{i=1}^N q(\boldsymbol{\theta}_{n,t}|\boldsymbol{\mu}_{i,t}, \mathbf{C}), \quad (144)$$

$$\Phi(\boldsymbol{\theta}_{n,t}) = q(\boldsymbol{\theta}_{n,t}|\boldsymbol{\mu}_{n,t}, \mathbf{C}), \quad (145)$$

4. Return all the pairs $\{\boldsymbol{\theta}_{n,t}, w_{n,t}\}$, and $\hat{Z} = \frac{1}{NT} \sum_{t=1}^T \sum_{n=1}^N w_{n,t}$.

of LAIS when $N = 1$, and $\{\boldsymbol{\mu}_t = \boldsymbol{\theta}_t\}$ i.e., all the samples $\{\boldsymbol{\theta}_t\}_{t=1}^T$ are generated by the unique MCMC chain with random walk proposal $\varphi(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}) = q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})$ with invariant density $P(\boldsymbol{\theta}|\mathbf{y})$. In this scenario, the two layers of LAIS are collapsed in a unique layer, so that $\{\boldsymbol{\mu}_t = \boldsymbol{\theta}_t\}$. Namely, no additional generation of samples are needed in the lower layer, and the samples generated in the upper layer (via MCMC) are recycled. Hence, the number of posterior evaluations is only T . The denominator for weights used in [81] is in Eq. (143), i.e., a temporal mixture as in [82]. The resulting estimator is

$$\hat{Z} = \frac{1}{T} \sum_{t=1}^T \frac{\pi(\boldsymbol{\theta}_t|\mathbf{y})}{\frac{1}{T} \sum_{k=1}^T \varphi(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})}, \quad \{\boldsymbol{\theta}_t\}_{t=1}^T \sim P(\boldsymbol{\theta}|\mathbf{y}) \text{ (via MCMC with a proposal } \varphi(\cdot|\cdot)\text{)}.$$

Relationship with KDE method. LAIS can be interpreted as an extension of the KDE method in Section 3, where the KDE function is also employed as a proposal density in the MIS scheme. Namely, the points used in Eq. (26), in LAIS they are drawn from the KDE function using the deterministic mixture procedure [51, 56, 57].

Compressed LAIS (CLAIS). Let us consider the T or N is large (i.e., either large chains or several parallel chains; or both). Since NT is large, the computation of the denominators Eqs. (142)- (143)- (144) can be expensive. A possible solution is to use a partitioning or clustering procedure [83] with $K \ll NT$ clusters considering the NT samples, and then employ as denominator the function

$$\Phi(\boldsymbol{\theta}) = \sum_{k=1}^K \bar{a}_k \mathcal{N}(\boldsymbol{\theta} | \bar{\boldsymbol{\mu}}_k, \mathbf{C}_k), \quad (146)$$

where $\bar{\boldsymbol{\mu}}_k$ represents the centroid of the k -th cluster, the normalized weight \bar{a}_k is proportional to the number of elements in the k -th cluster ($\sum_{k=1}^K \bar{a}_k = 1$), and $\mathbf{C}_k = \boldsymbol{\Sigma}_k + h\mathbf{I}$ with $\boldsymbol{\Sigma}_k$ the empirical covariance matrix of k -th cluster and $h > 0$.

Relationship with other methods using tempered posteriors. In the upper layer of LAIS, we can use non-tempered versions of the posterior, i.e., $P_n(\boldsymbol{\theta} | \mathbf{y}) = P(\boldsymbol{\theta} | \mathbf{y})$ for all n , or tempered versions of the posterior $P_n(\boldsymbol{\theta} | \mathbf{y}) = P(\boldsymbol{\theta} | \mathbf{y}, \beta_n) = \ell(\mathbf{y} | \boldsymbol{\theta})^{\beta_n} g(\boldsymbol{\theta})$. However, unlike in SS and/or power posterior methods, these samples are employed only as location parameters $\boldsymbol{\mu}_{n,t}$ of the proposal pdfs $q_{n,t}(\boldsymbol{\theta} | \boldsymbol{\mu}_{n,t}, \mathbf{C})$, and they are not included in the final estimators. Combining the tempered posteriors idea and the approach in [81], we could recycle $\boldsymbol{\theta}_{n,t} = \boldsymbol{\mu}_{n,t}$ and use $q_{n,t}(\boldsymbol{\theta} | \boldsymbol{\mu}_{n,t}) = \varphi_{n,t}(\boldsymbol{\theta} | \boldsymbol{\mu}_{n,t})$ where we denote as $\varphi_{n,t}$ the proposal pdfs employed in the MCMC chains. Another difference is that, in LAIS, the use of an “anti-tempered” posteriors with $\beta_n > 1$ is allowed and can be shown that is beneficial for the performance of the estimators (after the chains reach a good mixing) [84]. More generally, one can consider a time-varying $\beta_{n,t}$ (where t is the iteration of the n -th chain). In the first iterations, one could use $\beta_{n,t} < 1$ for fostering the exploration of the state space and helping the mixing of the chain. Then, in the last iterations, one could use $\beta_{n,t} > 1$ which increases the efficiency of the resulting IS estimators [84].

6 Vertical likelihood representations

In this section, we introduce a different approach based on Lebesgue representations of the integral expressing the marginal likelihood Z . First of all, we derive two one-dimensional integral representations of Z , and then we describe how it is possible to use these alternative representations by applying one-dimensional quadratures. However, the application of these quadrature rules is not straightforward. A possible final solution is the so-called nested sampling method.

6.1 Lebesgue representations of the marginal likelihood

6.1.1 First one-dimensional representation

The D_x -dimensional integral $Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$ can be turned into a one-dimensional integral using an extended space representation. Namely, we can write

$$Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} \quad (147)$$

$$= \int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} \int_0^{\ell(\mathbf{y}|\boldsymbol{\theta})} d\lambda \quad (\text{extended space representation}) \quad (148)$$

$$= \int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} \int_0^{\infty} \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}d\lambda \quad (149)$$

where $\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}$ is an indicator function which is 1 if $\lambda \in [0, \ell(\mathbf{y}|\boldsymbol{\theta})]$ and 0 otherwise. Switching the integration order, we obtain

$$Z = \int_0^{\infty} d\lambda \int_{\Theta} g(\boldsymbol{\theta})\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}d\boldsymbol{\theta} \quad (150)$$

$$= \int_0^{\infty} d\lambda \int_{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda} g(\boldsymbol{\theta})d\boldsymbol{\theta} \quad (151)$$

$$= \int_0^{\infty} Z(\lambda)d\lambda = \int_0^{\sup \ell(\mathbf{y}|\boldsymbol{\theta})} Z(\lambda)d\lambda, \quad (152)$$

where we have set

$$Z(\lambda) = \int_{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda} g(\boldsymbol{\theta})d\boldsymbol{\theta}. \quad (153)$$

In Eq. (152), we have also assumed that $\ell(\mathbf{y}|\boldsymbol{\theta})$ is bounded so the limit of integration is $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$. Below, we define several variables and sampling procedures required for the proper understanding of the nested sampling algorithm.

6.1.2 The survival function $Z(\lambda)$ and related sampling procedures

The function above $Z(\lambda) : \mathbb{R}^+ \rightarrow [0, 1]$ is the mass of the prior restricted to the set $\{\boldsymbol{\theta} : \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}$. Note also that

$$Z(\lambda) = \mathbb{P}(\lambda < \ell(\mathbf{y}|\boldsymbol{\theta})), \quad \text{where } \boldsymbol{\theta} \sim g(\boldsymbol{\theta}). \quad (154)$$

Moreover, we have that $Z(\lambda) \in [0, 1]$ with $Z(0) = 1$ and $Z(\lambda') = 0$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$, and it is also an non-increasing function. Therefore, $Z(\lambda)$ is a *survival function*, i.e.,

$$F(\lambda) = 1 - Z(\lambda) = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) < \lambda) = \mathbb{P}(\Lambda < \lambda), \quad (155)$$

is the cumulative distribution of the random variable $\Lambda = \ell(\mathbf{y}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$ [85, 2].

Sampling according to $F(\lambda) = 1 - Z(\lambda)$. Since $\Lambda = \ell(\mathbf{y}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$, the following procedure generates samples λ_n from $\frac{dF(\lambda)}{d\lambda}$:

1. Draw $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta})$, for $n = 1, \dots, N$.
2. Set $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$, for all $n = 1, \dots, N$.

Recalling the inversion method [85, Chapter 2], note also that the corresponding values

$$b_n = F(\lambda_n) \sim \mathcal{U}([0, 1]), \quad (156)$$

i.e., they are uniformly distributed in $[0, 1]$. Since $Z(\lambda) = 1 - F(\lambda)$, and since $V = 1 - U$ is also uniformly distributed $\mathcal{U}([0, 1])$ if $U \sim \mathcal{U}([0, 1])$, then

$$a_n = Z(\lambda_n) \sim \mathcal{U}([0, 1]). \quad (157)$$

In summary, finally we have that

$$\text{if } \boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}), \text{ and } \lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \sim F(\lambda) \quad \text{then} \quad a_n = Z(\lambda_n) \sim \mathcal{U}([0, 1]). \quad (158)$$

6.1.3 The truncated prior pdf $g(\boldsymbol{\theta}|\lambda)$ and other sampling procedures

Note that $Z(\lambda)$ is also the normalizing constant of the following truncated prior pdf

$$g(\boldsymbol{\theta}|\lambda) = \frac{1}{Z(\lambda)} \mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\} g(\boldsymbol{\theta}), \quad (159)$$

where $g(\boldsymbol{\theta}|0) = g(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta}|\lambda)$ for $\lambda > 0$. Two graphical examples of $g(\boldsymbol{\theta}|\lambda)$ and $Z(\lambda)$ are given in Figure 7.

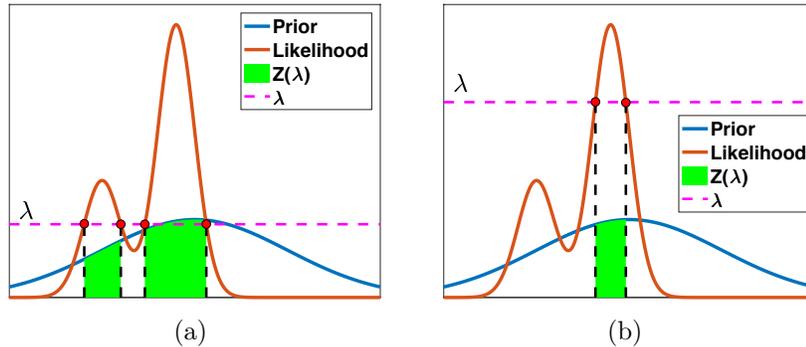


Figure 7: Two examples of the area below the truncated prior $g(\boldsymbol{\theta}|\lambda)$, i.e., the function $Z(\lambda)$. Note that in figure (b) the value of λ is greater than in figure (a), so that the area $Z(\lambda)$ decreases. If λ is bigger than the maximum of the likelihood function then $Z(\lambda) = 0$.

Sampling from $g(\boldsymbol{\theta}|\lambda)$ and $F(\lambda|\lambda_0)$. Given a fixed value $\lambda_0 \geq 0$, in order to generate samples from $g(\boldsymbol{\theta}|\lambda_0)$ one alternative is to use an MCMC procedure. However, in this case, the following acceptance-rejection procedure can be also employed [85]:

1. For $n = 1, \dots, N$:

- (a) Draw $\boldsymbol{\theta}' \sim g(\boldsymbol{\theta})$.
- (b) if $\ell(\mathbf{y}|\boldsymbol{\theta}') > \lambda_0$ then set $\boldsymbol{\theta}_n = \boldsymbol{\theta}'$ and $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}')$.
- (c) if $\ell(\mathbf{y}|\boldsymbol{\theta}') \leq \lambda_0$, then reject $\boldsymbol{\theta}'$ and repeat from step 1(a).

2. Return $\{\boldsymbol{\theta}_n\}_{n=1}^N$ and $\{\lambda_n\}_{n=1}^N$.

Observe that $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_0)$, for all $n = 1, \dots, N$, and the probability of accepting a generated sample $\boldsymbol{\theta}'$ is exactly $Z(\lambda)$. The values $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$ where $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_0)$, have the following *truncated* cumulative distribution

$$F(\lambda|\lambda_0) = \frac{F(\lambda) - F(\lambda_0)}{1 - F(\lambda_0)}, \quad \text{with } \lambda \geq \lambda_0, \quad (160)$$

i.e., we can write $\lambda_n \sim F(\lambda|\lambda_0)$.

6.1.4 Distribution of $a_n = Z(\lambda_n)$ and $\tilde{a}_n = \frac{a_n}{a_0}$ if $\lambda_n \sim F(\lambda|\lambda_0)$

Considering the values $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$ where $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_0)$, then $\lambda_n \sim F(\lambda|\lambda_0)$. Therefore, considering the values $a_0 = Z(\lambda_0) \leq 1$ and $a_n = Z(\lambda_n)$, with a similar argument used above in Eqs. (157)-(158) we can write

$$\begin{aligned} a_n &\sim \mathcal{U}([0, a_0]), \\ \tilde{a}_n &= \frac{a_n}{a_0} \sim \mathcal{U}([0, 1]), \quad \forall n = 1, \dots, N. \end{aligned}$$

In summary, with $a_0 = Z(\lambda_0)$, we have that

$$\text{if } \boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_0) \text{ and } \lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \sim F(\lambda|\lambda_0), \quad \text{then } Z(\lambda_n) \sim \mathcal{U}([0, a_0]), \quad (161)$$

and the ratio $\tilde{a}_n = \frac{a_n}{a_0} \sim \mathcal{U}([0, 1])$.

6.1.5 Distributions \tilde{a}_{\max}

Let us consider $\lambda_1, \dots, \lambda_n \sim F(\lambda|\lambda_0)$ and the minimum and maximum values

$$\lambda_{\min} = \min_n \lambda_n, \quad a_{\max} = Z(\lambda_{\min}), \quad \text{and} \quad \tilde{a}_{\max} = \frac{a_{\max}}{a_0} = \frac{Z(\lambda_{\min})}{Z(\lambda_0)}. \quad (162)$$

Let us recall $\tilde{a}_n = \frac{a_n}{a_0} \sim \mathcal{U}([0, 1])$. Then, note that \tilde{a}_{\max} is maximum of N uniform random variables

$$\tilde{a}_1, \dots, \tilde{a}_N \sim \mathcal{U}([0, 1]).$$

Then it is well-known that the cumulative distribution of the maximum value

$$\tilde{a}_{\max} = \max_n \tilde{a}_n \sim \mathcal{B}(N, 1),$$

is distributed according to a Beta distribution $\mathcal{B}(N, 1)$, i.e., $F_{\max}(\tilde{a}) = \tilde{a}^N$ and density $f_{\max}(\tilde{a}) = \frac{dF_{\max}(\tilde{a})}{d\tilde{a}} = N\tilde{a}^{N-1}$ [85, Section 2.3.6]. In summary, we have

$$\tilde{a}_{\max} = \frac{Z(\lambda_{\min})}{Z(\lambda_0)} \sim \mathcal{B}(N, 1), \text{ where } \lambda_{\min} = \min_n \lambda_n, \text{ and } \lambda_n \sim F(\lambda|\lambda_0). \quad (163)$$

This result is important for deriving the standard version of the nested sampling method, described in the next section. A summary of the relationships presented above is provided in Table 15.

Table 15: Summary of the relationships among the random variables introduced above.

Sections	Relationships
6.1.2	$Z(\lambda) = \mathbb{P}(\lambda < \ell(\mathbf{y} \boldsymbol{\theta}))$, and $F(\lambda) = 1 - Z(\lambda) = \mathbb{P}(\ell(\mathbf{y} \boldsymbol{\theta}) \leq \lambda)$, where $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$.
6.1.2	If $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta})$, we have $\lambda_n = \ell(\mathbf{y} \boldsymbol{\theta}_n) \sim F(\lambda)$ and $a_n = Z(\lambda_n) \sim \mathcal{U}([0, 1])$.
6.1.3 6.1.4	If $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta} \lambda_0)$, we have $\lambda_n = \ell(\mathbf{y} \boldsymbol{\theta}_n) \sim F(\lambda \lambda_0)$ and $a_n = Z(\lambda_n) \sim \mathcal{U}([0, a_0])$, with $a_0 = Z(\lambda_0)$. Moreover, $\tilde{a}_n = \frac{a_n}{a_0} \sim \mathcal{U}([0, 1])$.
6.1.5	If $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta} \lambda_0)$, we have $\lambda_n = \ell(\mathbf{y} \boldsymbol{\theta}_n) \sim F(\lambda \lambda_0)$ and $\tilde{a}_{\max} = \frac{Z(\lambda_{\min})}{Z(\lambda_0)} \sim \mathcal{B}(N, 1)$, where $\lambda_{\min} = \min_n \lambda_n$. Note also that $\tilde{a}_{\max} = \max_n \tilde{a}_n$.

6.1.6 Second one-dimensional representation

Now let consider a specific *area* value $a = Z(\lambda)$. The inverse function

$$\Psi(a) = Z^{-1}(a) = \sup\{\lambda : Z(\lambda) > a\}, \quad (164)$$

is also non-increasing. Note that $Z(\lambda) > a$ if and only if $\lambda < \Psi(a)$. Then, we can write

$$\begin{aligned}
Z &= \int_0^\infty Z(\lambda) d\lambda \\
&= \int_0^\infty d\lambda \int_0^1 \mathbb{I}\{a < Z(\lambda)\} da && \text{(again the extended space "trick")} \\
&= \int_0^1 da \int_0^\infty \mathbb{I}\{u < Z(\lambda)\} d\lambda && \text{(switching the integration order)} \\
&= \int_0^1 da \int_0^\infty \mathbb{I}\{\lambda < \Psi(a)\} d\lambda && \text{(using } Z(\lambda) > a \iff \lambda < \Psi(a)\text{)} \\
&= \int_0^1 \Psi(a) da.
\end{aligned} \tag{165}$$

6.1.7 Summary of the one-dimensional representations

Thus, finally we have obtained two one-dimensional integrals for expressing the Bayesian evidence Z ,

$$Z = \int_0^{\sup \ell(\mathbf{y}|\boldsymbol{\theta})} Z(\lambda) d\lambda = \int_0^1 \Psi(a) da. \tag{166}$$

Now that we have expressed the quantity Z as an integral of a function over \mathbb{R} , we could think of applying simple quadrature: choose a grid of points in $[0, \sup \ell(\mathbf{y}|\boldsymbol{\theta})]$ ($\lambda_i > \lambda_{i-1}$) or in $[0, 1]$ ($a_i > a_{i-1}$), evaluate $Z(\lambda)$ or $\Psi(a)$ and use the quadrature formulas

$$\widehat{Z} = \sum_{i=1}^I (\lambda_i - \lambda_{i-1}) Z(\lambda_i), \text{ or} \tag{167}$$

$$\widehat{Z} = \sum_{i=1}^I (a_i - a_{i-1}) \Psi(a_i). \tag{168}$$

However, this simple approach is not desirable since (i) the functions $Z(\lambda)$ and $\Psi(a)$ are intractable in most cases and (ii) they change much more rapidly over their domains than does $\pi(\boldsymbol{\theta}|\mathbf{y}) = \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})$, hence the quadrature approximation can have very bad performance, unless the grid of points is chosen with extreme care. Table 16 summarizes the one-dimensional expression for $\log Z$ and Z contained in this work. Clearly, in all of them, the integrand function depends, explicitly or implicitly, on the variable $\boldsymbol{\theta}$.

6.2 Nested Sampling

Nested sampling is a technique for estimating the marginal likelihood that exploits the second identity in (166) [28, 86, 26]. Nested Sampling estimates Z by a quadrature using nodes (in *decreasing* order),

$$0 < a_{\max}^{(I)} < \dots < a_{\max}^{(1)} < 1$$

Table 16: One-dimensional integrals for $\log Z$ and Z . Note that, in all cases, the integrand function contains the dependence on $\boldsymbol{\theta}$.

Method	Expression	Equations
path sampling	$\log Z = \int_0^1 \frac{1}{Z(\beta)} \frac{\partial}{\partial \beta} \left(\int_{\Theta} \pi(\boldsymbol{\theta} \mathbf{y}, \beta) d\boldsymbol{\theta} \right) d\beta$	(98)
power-posteriors	$\log Z = \int_0^1 \mathbb{E}_{P(\boldsymbol{\theta} \mathbf{y}, \beta)} [\log \ell(\mathbf{y} \boldsymbol{\theta})] d\beta$	(103)
vertical representation-1	$Z = \int_0^{\sup \ell(\mathbf{y} \boldsymbol{\theta})} Z(\lambda) d\lambda$	(152)-(153)
vertical representation-2	$Z = \int_0^1 \Psi(a) da$	(165)

and the quadrature formula

$$\widehat{Z} = \sum_{i=1}^I (a_{\max}^{(i-1)} - a_{\max}^{(i)}) \Psi(a_{\max}^{(i)}) = \sum_{i=1}^I (a_{\max}^{(i-1)} - a_{\max}^{(i)}) \lambda_{\min}^{(i)}, \quad (169)$$

with $a_{\max}^{(0)} = 1$. We have to specify the grid points $a_{\max}^{(i)}$'s (possibly well-located, with a suitable strategy) and the corresponding values $\lambda_{\min}^{(i)} = \Psi(a_{\max}^{(i)})$. Recall that the function $\Psi(a)$, and its inverse $a = \Psi^{-1}(\lambda) = Z(\lambda)$, are generally intractable, so that it is not even possible to evaluate $\Psi(a)$ at a grid of chosen $a_{\max}^{(i)}$'s.

Remark 19. *The nested sampling algorithm works in the other way around: it suitably selects the ordinates $\lambda_{\min}^{(i)}$'s and find some approximations \widehat{a}_i 's of the corresponding values $a_{\max}^{(i)} = Z(\lambda_{\min}^{(i)})$. This is possible since the distribution of $a_{\max}^{(i)}$ is known (see Section 6.1.5).*

6.2.1 Choice of $\lambda_{\min}^{(i)}$ and $a_{\max}^{(i)}$ in nested sampling

Nested sampling employs an iterative procedure in order to generate an *increasing* sequence of likelihood ordinates $\lambda_{\min}^{(i)}$, $i = 1, \dots, I$, such that

$$\lambda_{\min}^{(1)} < \lambda_{\min}^{(2)} < \lambda_{\min}^{(3)} \dots < \lambda_{\min}^{(I)}. \quad (170)$$

The details of the algorithm is given in Table 17 and it is based on the sampling of the truncated prior pdf $g(\boldsymbol{\theta}|\lambda_{\min}^{(i-1)})$ (see Sections from 6.1.2 to 6.1.5), where i denotes the iteration index. The nested sampling procedure is explained below:

- At the first iteration ($i = 1$), we set $\lambda_{\min}^{(0)} = 0$ and $a_{\max}^{(0)} = Z(\lambda_{\min}^{(0)}) = 1$. Then, N samples are drawn from the prior $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_{\min}^{(0)}) = g(\boldsymbol{\theta})$ obtaining a cloud $\mathcal{P} = \{\boldsymbol{\theta}_n\}_{n=1}^N$ and then set $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$, i.e., $\{\lambda_n\}_{n=1}^N \sim F(\lambda)$ as shown in Section 6.1.2. Thus, the first ordinate is chosen as

$$\lambda_{\min}^{(1)} = \min_n \lambda_n = \min_n \ell(\mathbf{y}|\boldsymbol{\theta}_n) = \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\boldsymbol{\theta}).$$

Since $\{\lambda_n\}_{n=1}^N \sim F(\lambda)$, using the result in Eq. (163), we have that

$$\tilde{a}_{\max}^{(1)} = \frac{a_{\max}^{(1)}}{a_{\max}^{(0)}} = \frac{Z(\lambda_{\min}^{(1)})}{Z(\lambda_{\min}^{(0)})} \sim \mathcal{B}(N, 1).$$

Since $a_{\max}^{(0)} = Z(\lambda_{\min}^{(0)}) = 1$, then $\tilde{a}_{\max}^{(1)} = a_{\max}^{(1)} \sim \mathcal{B}(N, 1)$. The corresponding $\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$ is also removed from \mathcal{P} , i.e., $\mathcal{P} = \mathcal{P} \setminus \{\boldsymbol{\theta}^*\}$ (now $|\mathcal{P}| = N - 1$).

- At a generic i -th iteration ($i \geq 2$), a unique additional sample $\boldsymbol{\theta}'$ is drawn from the truncated prior $g(\boldsymbol{\theta}|\lambda_{\min}^{(i-1)})$ and added to the current cloud of samples, i.e., $\mathcal{P} = \mathcal{P} \cup \boldsymbol{\theta}'$ (now again $|\mathcal{P}| = N$). First of all, note that the value $\lambda' = \lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}')$ is distributed as $F(\lambda|\lambda_{\min}^{(i-1)})$ (see Section 6.1.3). More precisely, note that all the N ordinate values

$$\{\lambda_n\}_{n=1}^N = \ell(\mathbf{y}|\mathcal{P}) = \{\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \text{ for all } \boldsymbol{\theta}_n \in \mathcal{P}\}$$

are distributed as $F(\lambda|\lambda_{\min}^{(i-1)})$, i.e., $\{\lambda_n\}_{n=1}^N \sim F(\lambda|\lambda_{\min}^{(i-1)})$. This is due to how the population \mathcal{P} has been built in the previous iterations. Then, we choose the new minimum value as

$$\lambda_{\min}^{(i)} = \min_n \lambda_n = \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\mathcal{P}).$$

Moreover, since $\lambda_{\min}^{(i)}$ is the minimum value of $\{\lambda_1, \dots, \lambda_N\} \sim F(\lambda|\lambda_{\min}^{(i-1)})$, in Section 6.1.5 we have seen that

$$\tilde{a}_{\max}^{(i)} = \frac{a_{\max}^{(i)}}{a_{\max}^{(i-1)}} = \frac{Z(\lambda_{\min}^{(i)})}{Z(\lambda_{\min}^{(i-1)})} \sim \mathcal{B}(N, 1), \quad (171)$$

where we have used Eq. (163). We remove again the corresponding sample $\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$, i.e., we set $\mathcal{P} = \mathcal{P} \setminus \{\boldsymbol{\theta}^*\}$ and the procedure is repeated. Note that we have also found the recursion among the following random variables,

$$a_{\max}^{(i)} = \tilde{a}_{\max}^{(i)} a_{\max}^{(i-1)}, \quad (172)$$

for $i = 1, \dots, I$ and $a_{\max}^{(0)} = 1$.

- The random value $\tilde{a}_{\max}^{(i)}$ could be estimated and replaced with the expected value of the Beta distribution $\mathcal{B}(N, 1)$, i.e.,

$$\tilde{a}_{\max}^{(i)} \approx \hat{a}_1 = \frac{N}{N+1} \approx \exp\left(-\frac{1}{N}\right). \quad (173)$$

where $\mathbb{E}[\mathcal{B}(N, 1)] = \frac{N}{N+1}$, and $\exp\left(-\frac{1}{N}\right)$ becomes a very good approximation as N grows. In that case, the recursion above becomes

$$a_{\max}^{(i)} \approx \exp\left(-\frac{1}{N}\right) a_{\max}^{(i-1)} = \exp\left(-\frac{i}{N}\right). \quad (174)$$

Then, denoting $\hat{a}_i = \exp\left(-\frac{i}{N}\right)$, we can use \hat{a}_i as an approximation of $a_{\max}^{(i)}$.

Remark 20. *The intuition behind the iterative approach above is to accumulate more ordinates λ_i close to the $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$. They are also more dense around $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$. Moreover, using this scheme, we can employ $\hat{a}_i = \exp\left(-\frac{i}{N}\right)$ as an approximation of $a_{\max}^{(i)}$.*

Remark 21. *An implicit optimization of the likelihood function is performed in the nested sampling algorithm. All population of $\lambda_i \in \mathcal{P}$ approaches the value $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$.*

Table 17: The standard Nested Sampling procedure.

1. Choose N and set $\hat{a}_0 = 1$.
2. Draw $\{\boldsymbol{\theta}_n\}_{n=1}^N \sim g(\boldsymbol{\theta})$ and define the set $\mathcal{P} = \{\boldsymbol{\theta}_n\}_{n=1}^N$. Let us also define the notation
$$\ell(\mathbf{y}|\mathcal{P}) = \{\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \text{ for all } \boldsymbol{\theta}_n \in \mathcal{P}\}, \quad (175)$$
3. Set $\lambda_{\min}^{(1)} = \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\boldsymbol{\theta})$ and $\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\boldsymbol{\theta})$.
4. Set $\mathcal{P} = \mathcal{P} \setminus \{\boldsymbol{\theta}^*\}$, i.e., eliminate $\boldsymbol{\theta}^*$ from \mathcal{P} .
5. Find an approximation \hat{a}_1 of $a_{\max}^{(1)} = Z(\lambda_{\min}^{(1)})$. One usual choice is $\hat{a}_1 = \exp\left(-\frac{1}{N}\right)$.
6. For $i = 2, \dots, I$:
 - (a) Draw $\boldsymbol{\theta}' \sim g(\boldsymbol{\theta}|\lambda_{\min}^{(i-1)})$ and add to the current cloud of samples, i.e., $\mathcal{P} = \mathcal{P} \cup \boldsymbol{\theta}'$.
 - (b) Set $\lambda_{\min}^{(i)} = \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\boldsymbol{\theta})$ and $\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\boldsymbol{\theta})$.
 - (c) Set $\mathcal{P} = \mathcal{P} \setminus \{\boldsymbol{\theta}^*\}$.
 - (d) Find an approximation \hat{a}_i of $a_{\max}^{(i)} = Z(\lambda_{\min}^{(i)})$. One usual choice is
$$\hat{a}_i = \exp\left(-\frac{i}{N}\right), \quad (176)$$

The rationale behind this choice is explained in the section above.

7. Return

$$\hat{Z} = \sum_{i=1}^I (\hat{a}_{i-1} - \hat{a}_i) \lambda_{\min}^{(i)} = \sum_{i=1}^I (e^{-\frac{i-1}{N}} - e^{-\frac{i}{N}}) \lambda_{\min}^{(i)}. \quad (177)$$

6.2.2 Further considerations

Perhaps, the most critical task of the nested sampling implementation consists in drawing from the truncated priors. For this purpose, one can use a rejection sampling or an MCMC scheme. In

the first case, we sample from the prior and then accept only the samples $\boldsymbol{\theta}'$ such that $\ell(\mathbf{y}|\boldsymbol{\theta}') > \lambda$. However, as λ grows, its performance deteriorates since the acceptance probability gets smaller and smaller. The MCMC algorithms could also have poor performance due to the sample correlation, specially when the support of the constrained prior is formed by disjoint regions or distant modes [86]. Moreover, in the derivation of the standard nested sampling method we have considered different approximations. First of all, for each likelihood value λ_i , its corresponding $a_i = \Psi^{-1}(\lambda_i)$ is approximated by replacing the expected value of a Beta random variable within a recursion involving a_i (Eq. (172)). Then this expected value is again approximated with an exponential function in Eq. (173). This step could be avoided, keeping directly $\frac{N}{N+1}$. The simplicity of the final formula $\hat{a}_i = \exp(-\frac{i}{N})$ is perhaps the reason of using the approximation $\frac{N}{N+1} \approx \exp(-\frac{1}{N})$. A further approximation $\mathbb{E}[a_{\max}^{(i)}] \approx \mathbb{E}[\tilde{a}_{\max}^{(i)}]\mathbb{E}[a_{\max}^{(i-1)}]$ is also implicitly applied in (174). Additionally, if an MCMC method is run for sampling from the constrained prior, also the likelihood values λ_i are in some sense approximated due to the possible burn-in period of the chain.

6.2.3 Generalized Importance Sampling based on vertical representations

Let us recall the estimator IS vers-2 with proposal density $\bar{q}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})$,

$$\hat{Z} = \sum_{n=1}^N \bar{\rho}_n \ell(\mathbf{y}|\boldsymbol{\theta}_n), \quad \{\boldsymbol{\theta}_n\}_{n=1}^N \sim \bar{q}(\boldsymbol{\theta}), \quad (178)$$

where $\rho_n = \frac{g(\boldsymbol{\theta}_n)}{q(\boldsymbol{\theta}_n)}$ and $\bar{\rho}_n = \frac{\rho_n}{\sum_{n=1}^N \rho_n}$. In [26], the authors consider the use of the following proposal pdf

$$\bar{q}_w(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})W(\ell(\mathbf{y}|\boldsymbol{\theta}))}{Z_w} \propto q_w(\boldsymbol{\theta}) = g(\boldsymbol{\theta})W(\ell(\mathbf{y}|\boldsymbol{\theta})), \quad (179)$$

where the function $W(\lambda) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is defined by the user. Using $\bar{q}_w(\boldsymbol{\theta})$ leads to the weights of the form

$$\rho_n = \frac{g(\boldsymbol{\theta}_n)}{q_w(\boldsymbol{\theta}_n)} = \frac{1}{W(\ell(\mathbf{y}|\boldsymbol{\theta}_n))}, \quad \boldsymbol{\theta}_n \sim \bar{q}_w(\boldsymbol{\theta}). \quad (180)$$

Note that choosing $W(\lambda) = \lambda$ we have $W(\ell(\mathbf{y}|\boldsymbol{\theta})) = \ell(\mathbf{y}|\boldsymbol{\theta})$, and $\bar{q}_w(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y})$, recovering the harmonic mean estimator. With $W(\lambda) = \lambda^\beta$, we have $W(\ell(\mathbf{y}|\boldsymbol{\theta})) = \ell(\mathbf{y}|\boldsymbol{\theta})^\beta$ and $\bar{q}_w(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^\beta}{Z(\beta)}$, recovering the method in Section 4.3.2 that uses a power posterior as a proposal pdf. Nested sampling seems that can be also included in this framework [26].

7 On the marginal likelihood approach and other strategies

In this section, we examine the marginal likelihood approach to Bayesian model selection and compare it to other strategies such as the well-known *posterior predictive check* approach.

7.1 Dependence on the prior and related discussion

The marginal likelihood approach for model selection and hypothesis testing naturally appears as a consequence of the application of Bayes' theorem to derive posterior model probabilities $p(\mathcal{M}_m|\mathbf{y}) \propto p_m Z_m$. Under the assumption that one of \mathcal{M}_m is the true generating model, the Bayes factor will choose the correct model as the number of data grows, $D_y \rightarrow \infty$ [87]. We can also apply the posterior model probabilities $p(\mathcal{M}_m|\mathbf{y})$ to combine inferences across models, a setting called Bayesian model averaging [19, 20].

7.1.1 Dependence on the prior

In Section 2.2, we have seen the marginal likelihood Z contains intrinsically a penalization for the model complexity. This penalization is related to the choice of the prior and its “overlap” with likelihood function. Indeed, $Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$ is by definition a continuous mixture of the likelihood values weighted according to the prior. In this sense, depending on the choice of the prior, the evidence Z can take any possible value in the interval $[\ell(\mathbf{y}|\boldsymbol{\theta}_{\min}), \ell(\mathbf{y}|\boldsymbol{\theta}_{\max})]$ (see Section 2.2, for more details). Hence, the marginal likelihood even with strong data (unlike the posterior density) is highly sensitivity to the choice of prior density. See also the examples in the Supplementary Material.

Improper priors. The use of improper priors, $\int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, is allowed when $\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, since the corresponding posteriors are proper. However, this is an issue for the model selection with Z . Indeed, the prior $g(\boldsymbol{\theta}) = c h(\boldsymbol{\theta})$ is not completely specified, since $c > 0$ is arbitrary. Some possible solutions are given in Section 7.2.

Generally, the use of more diffuse (proper) priors provides smaller values of Z . Therefore, different choices of the priors can yield different selected models. For this fact, some authors criticize the use of evidence Z for model comparison.

7.1.2 Safe scenarios for fair comparisons

In a Bayesian framework, the best scenario is clearly when the practitioners and/or researchers have strong beliefs that can be translated into informative priors. Hence, in this setting, the priors truly encode some relevant information about the inference problem. When this additional information is not available, different strategies could be considered. We consider as a safe scenario for comparing different models, a scenario where the choice of the priors is *virtually* not favoring any of the models. Below and in Sections 7.2 and 7.4, we describe some interesting scenarios and some possible solutions for reducing, in some way, the dependence of the model comparison on the choice of the priors.

Same priors. Generally, we are interested in comparing two or more models. The use of the same (even improper) priors is possible when the models have the same parameters (and hence also share the same support space). With this choice, the resulting comparison seems fair and reasonable. However, this scenario is very restricted in practice. An example is when we have

nested models. As noted in [87, Sect. 5.3], in the context of testing hypothesis, some authors have considered improper priors on nuisance parameters that appear on both null and alternative hypothesis. Since the nuisance parameters appear on both models, the multiplicative constants cancel out in the Bayes factor.

Likelihood-based priors. When $\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, we can build a prior based on the data and the observation model. For instance, we can choose $g_{\text{like}}(\boldsymbol{\theta}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})}{\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}$, then the marginal likelihood is

$$Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g_{\text{like}}(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{\int_{\Theta} \ell^2(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \quad (181)$$

This idea is connected to *posterior predictive approach*, described in Section 7.4. Indeed, the marginal likelihood above can be written as $Z = E_{P(\boldsymbol{\theta}|\mathbf{y})}[\ell(\mathbf{y}|\boldsymbol{\theta})] = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ when $g(\boldsymbol{\theta}) = 1$. Less informative likelihood-based priors can be constructed using a tempering effect with a parameter $0 < \beta \leq 1$ or considering only a subset of data \mathbf{y}_{sub} . For instance, when $\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})^\beta d\boldsymbol{\theta} < \infty$ or $\int_{\Theta} \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, then we can choose $g_{\text{like}}(\boldsymbol{\theta}) \propto \ell(\mathbf{y}|\boldsymbol{\theta})^\beta$ or $g_{\text{like}}(\boldsymbol{\theta}) \propto \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})$, the marginal likelihood is

$$Z = \frac{\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta+1}d\boldsymbol{\theta}}{\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})^\beta d\boldsymbol{\theta}}, \quad \text{or} \quad Z = \frac{\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})\ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Theta} \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \quad (182)$$

This is also the key idea underlying the partial and intrinsic Bayes factors described in the next section.

7.2 Bayes factors with improper priors

So far we have considered proper priors, i.e., $\int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$. The use of improper priors is common in Bayesian inference to represent weak prior information. Consider $g(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta})$ where $h(\boldsymbol{\theta})$ is a non-negative function whose integral over the state space does not converge, $\int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\Theta} h(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$. In that case, $g(\boldsymbol{\theta})$ is not completely specified. Indeed, we can have different definitions $g(\boldsymbol{\theta}) = ch(\boldsymbol{\theta})$ where $c > 0$ is (the inverse of) the “normalizing” constant, not uniquely determinate since c formally does not exist. Regarding the parameter inference and posterior definition, the use of improper priors poses no problems as long as $\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, indeed

$$\begin{aligned} P(\boldsymbol{\theta}|\mathbf{y}) &= \frac{1}{Z}\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})ch(\boldsymbol{\theta})}{\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})ch(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}}, \\ &= \frac{1}{Z_h}\ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta}) \end{aligned} \quad (183)$$

where $Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$, $Z_h = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $Z = cZ_h$. Note that the unspecified constant $c > 0$ is canceled out, so that the posterior $P(\boldsymbol{\theta}|\mathbf{y})$ is well-defined even with an improper prior if $\int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$. However, the issue is not solved when we compare different models,

since $Z = cZ_h$ depends on c . For instance, the Bayes factors depend on the undetermined constants $c_1, c_2 > 0$ [88],

$$\text{BF}(\mathbf{y}) = \frac{c_1 \int_{\Theta_1} \ell_1(\mathbf{y}|\boldsymbol{\theta}_1)h_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{c_2 \int_{\Theta_2} \ell_2(\mathbf{y}|\boldsymbol{\theta}_2)h_2(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2} = \frac{Z_1}{Z_2} = \frac{c_1 Z_{h_1}}{c_2 Z_{h_2}}, \quad (184)$$

so that different choices of c_1, c_2 provide different preferable models. There exists various approaches for dealing with this issue. Below we describe some relevant ones.

Partial Bayes Factors. The idea behind the partial Bayes factors consists of using a subset of data to build proper priors and, jointly with the remaining data, they are used to calculate the Bayes factors. This is related to the likelihood-based prior approach, described above. The method starts by dividing the data in two subsets, $\mathbf{y} = (\mathbf{y}_{\text{train}}, \mathbf{y}_{\text{test}})$. The first subset $\mathbf{y}_{\text{train}}$ is used to obtain partial posterior distributions,

$$\bar{g}_m(\boldsymbol{\theta}_m|\mathbf{y}_{\text{train}}) = \frac{c_m}{Z_{\text{train}}^{(m)}} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}_m)h_m(\boldsymbol{\theta}_m), \quad (185)$$

using the improper priors. The partial posterior $\bar{g}_m(\boldsymbol{\theta}_m|\mathbf{y}_{\text{train}})$ is then employed as prior. Note that

$$Z_{\text{train}}^{(m)} = c_m \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}_m)h_m(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m.$$

Recall that the complete posterior of m -th model is

$$P_m(\boldsymbol{\theta}|\mathbf{y}) = P_m(\boldsymbol{\theta}|\mathbf{y}_{\text{test}}, \mathbf{y}_{\text{train}}) = \frac{c_m}{Z_m} \ell_m(\mathbf{y}|\boldsymbol{\theta}_m)h_m(\boldsymbol{\theta}_m), \quad (186)$$

where

$$Z_m = c_m \int_{\Theta_m} \ell_m(\mathbf{y}|\boldsymbol{\theta}_m)h_m(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m.$$

Note that $Z_{\text{train}}^{(m)}$ and Z_m both depend on the unspecified constant c_m . Considering the conditional likelihood $\ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}_m, \mathbf{y}_{\text{train}})$ of the remaining data \mathbf{y}_{test} ,⁵ we can study another posterior of \mathbf{y}_{test} , that

$$P_{\text{test}}^{(m)}(\boldsymbol{\theta}|\mathbf{y}_{\text{test}}) = \frac{1}{Z_{\text{test}|\text{train}}^{(m)}} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}_m, \mathbf{y}_{\text{train}})\bar{g}_m(\boldsymbol{\theta}_m|\mathbf{y}_{\text{train}}), \quad (187)$$

⁵In case of conditional independence of the data given $\boldsymbol{\theta}$, we have $\ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}_m, \mathbf{y}_{\text{train}}) = \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}_m)$.

where $\bar{g}_m(\boldsymbol{\theta}_m|\mathbf{y}_{\text{train}})$ in (185) plays the role of a prior pdf, and

$$\begin{aligned}
Z_{\text{test}|\text{train}}^{(m)} &= \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}_m, \mathbf{y}_{\text{train}}) \bar{g}_m(\boldsymbol{\theta}_m|\mathbf{y}_{\text{train}}) d\boldsymbol{\theta}_m, \\
&= \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}_m, \mathbf{y}_{\text{train}}) \frac{c_m}{Z_{\text{train}}^{(m)}} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}_m) h_m(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m, \\
&= \frac{c_m}{Z_{\text{train}}^{(m)}} \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}_m, \mathbf{y}_{\text{train}}) \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}_m) h_m(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m, \\
&= \frac{c_m}{Z_{\text{train}}^{(m)}} \int_{\Theta_m} \ell_m(\mathbf{y}|\boldsymbol{\theta}_m) h_m(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m, \\
&= \frac{Z_m}{Z_{\text{train}}^{(m)}}.
\end{aligned}$$

Thus, $Z_{\text{test}|\text{train}}^{(m)}$ does not depend on c_m . Therefore, considering the partial posteriors $\bar{g}_m(\boldsymbol{\theta}_m|\mathbf{y}_{\text{train}})$ as proper priors, we can define the following *partial* Bayes factor

$$\begin{aligned}
\text{BF}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}}) &= \frac{Z_{\text{test}|\text{train}}^{(1)}}{Z_{\text{test}|\text{train}}^{(2)}} = \frac{\frac{Z_1}{Z_{\text{train}}^{(1)}}}{\frac{Z_2}{Z_{\text{train}}^{(2)}}}, \\
&= \frac{\frac{Z_1}{Z_2}}{\frac{Z_{\text{train}}^{(1)}}{Z_{\text{train}}^{(2)}}} = \frac{\text{BF}(\mathbf{y})}{\text{BF}(\mathbf{y}_{\text{train}})}. \quad (\text{“Bayes law for Bayes Factors”}). \quad (188)
\end{aligned}$$

Therefore, one can approximate firstly $\text{BF}(\mathbf{y}_{\text{train}})$, secondly $\text{BF}(\mathbf{y})$ and then compare the model using the partial Bayes factor $\text{BF}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}})$.

Remark 22. *The trick here consists in computing two normalizing constants for each model, instead of only one. The first normalizing constant is used for building an auxiliary proper prior, depending on $\mathbf{y}_{\text{train}}$. The difference with the likelihood-based prior approach in previous section is that $\mathbf{y}_{\text{train}}$ is used only once (in the auxiliary proper prior).*

A training dataset $\mathbf{y}_{\text{train}}$ is proper if $\int_{\Theta_m} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}_m) h_m(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m < \infty$ for all models, and it is called *minimal* if is proper and no subset of $\mathbf{y}_{\text{train}}$ is proper. If we use actually proper prior densities, the minimal training dataset is the empty set and the fractional Bayes factor reduces to the classical Bayes factor. However, the main drawback of the partial Bayes factor approach is the dependence on the choice of $\mathbf{y}_{\text{train}}$ (which could affect the selection of the model). The authors suggest finding the *minimal* suitable training set $\mathbf{y}_{\text{train}}$, but this task is not straightforward. Two alternatives in the literature have been proposed, the fractional Bayes factors and the intrinsic Bayes factors.

Fractional Bayes Factors [89]. Instead of using a training data, it is possible to use power posteriors, i.e.,

$$\text{FBF}(\mathbf{y}) = \frac{\text{BF}(\mathbf{y})}{\text{BF}(\mathbf{y}|\beta)}, \quad (189)$$

where the denominator is

$$\text{BF}(\mathbf{y}|\beta) = \frac{\int_{\Theta_1} \ell_1(\mathbf{y}|\boldsymbol{\theta}_1)^\beta g_1(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int_{\Theta_2} \ell_2(\mathbf{y}|\boldsymbol{\theta}_2)^\beta g_2(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2} = \frac{c_1 \int_{\Theta_1} \ell_1(\mathbf{y}|\boldsymbol{\theta}_1)^\beta h_1(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{c_2 \int_{\Theta_2} \ell_2(\mathbf{y}|\boldsymbol{\theta}_2)^\beta h_2(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2}. \quad (190)$$

with $0 < \beta < 1$, and $\text{BF}(\mathbf{y}|1) = \text{BF}(\mathbf{y})$. Note that the value $\beta = 0$ is not admissible since $\int_{\Theta_m} h_m(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m = \infty$ for $m = 1, 2$. Again, since both $\text{BF}(\mathbf{y})$ and $\text{BF}(\mathbf{y}|\beta)$ depend on the ratio $\frac{c_1}{c_2}$, the fractional Bayes factor $\text{FBF}(\mathbf{y})$ is independent on c_1 and c_2 by definition.

Intrinsic Bayes factors [90]. The partial Bayes factor (188) will depend on the choice of (minimal) training set $\mathbf{y}_{\text{train}}$. These authors solve the problem of choosing the training sample by averaging the partial Bayes factor over all possible minimal training sets. They suggest using the arithmetic mean, leading to the *arithmetic* intrinsic Bayes factor, or the geometric mean, leading to the *geometric* intrinsic Bayes factor.

7.3 Marginal likelihood as a prior predictive approach

Due to the definition of the marginal likelihood $Z = E_g[\ell(\mathbf{y}|\boldsymbol{\theta})] = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is also called or related to the so-called *prior predictive approach*. As in the Approximate Bayesian Computation (ABC) [91], the idea is that we can generate artificial data $\tilde{\mathbf{y}}_{i,m}$, $i = 1, \dots, L$ from each m -th model with the following procedure: (a) draw $\boldsymbol{\theta}_{i,m}$ from the m -th prior, $g_m(\boldsymbol{\theta})$ and $\tilde{\mathbf{y}}_{i,m}$ from the m -th likelihood $\ell_m(\mathbf{y}|\boldsymbol{\theta}_{i,m})$. Given each set of fake data $\mathcal{S}_m = \{\tilde{\mathbf{y}}_{i,m}\}_{i=1}^L$, we can use different classical hypothesis testing techniques for finding the set \mathcal{S}_m closest to the true data \mathbf{y} (for instance, based on p -values). Another possibility, we could approximate the value $Z_m = p_m(\mathbf{y})$ applying kernel density estimation \hat{p}_m to each set \mathcal{S}_m .

In the next section, we describe the posterior predictive approach, which consider the expected value of likelihood evaluated in a generic $\tilde{\mathbf{y}}$ with respect to (w.r.t.) the posterior $P(\boldsymbol{\theta}|\mathbf{y})$, instead of w.r.t. the prior $g(\boldsymbol{\theta})$. The posterior predictive idea can be considered an alternative model selection approach w.r.t. the marginal likelihood approach, which includes several well-known model selection schemes.

7.4 Other ways of model selection: the posterior predictive approach

The marginal likelihood approach is not the unique approach for model selection in Bayesian statistics. Here, we discuss some alternatives which are based on the concept of prediction.

After fitting a Bayesian model, a popular approach for model checking (i.e. assessing the adequacy of the model fit to the data) consists in measuring its predictive accuracy [92, Chapter 6][93]. Hence, a key quantity in these approaches is the posterior predictive distribution of generic different data $\tilde{\mathbf{y}}$ given \mathbf{y} ,

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = E_{P(\boldsymbol{\theta}|\mathbf{y})}[\ell(\tilde{\mathbf{y}}|\boldsymbol{\theta})] = \int_{\Theta} \ell(\tilde{\mathbf{y}}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \quad (191)$$

Considering $\tilde{\mathbf{y}} = \mathbf{y}$, note that exists a clear connection with likelihood-based priors described in Section 7.1.

Remark 23. *The posterior predictive distribution in (191) is an expectation w.r.t. the posterior, which is robust to the prior selection with informative data, unlike the marginal likelihood. Therefore, this approach is less affected by the prior choice.*

Note that we can consider posterior predictive distributions $p(\tilde{\mathbf{y}}|\mathbf{y})$ for vectors $\tilde{\mathbf{y}}$ smaller than \mathbf{y} (i.e., with less components). The *posterior predictive checking* is based on the main idea of considering some simulated data $\tilde{\mathbf{y}}_i \sim p(\tilde{\mathbf{y}}|\mathbf{y})$, with $i = 1, \dots, L$, and comparing with the observed data \mathbf{y} . After obtaining a set of fake data $\{\tilde{\mathbf{y}}_i\}_{i=1}^L$, we have to measure the discrepancy between the true observed data \mathbf{y} and the set $\{\tilde{\mathbf{y}}_i\}_{i=1}^L$. This comparison can be made with test quantities and graphical checks (e.g., posterior predictive p-values).

Alternatively, different measures of predictive accuracy can be employed. An example, is the *expected log pointwise predictive density* (ELPD) [94]. Let recall that $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$, and define as $\bar{y} \in \mathbb{R}$ any alternative scalar data. Considering M alternative scalar data \bar{y}_i with density $p_{\text{true}}(\bar{y}_i)$, the ELPD is defined as

$$\begin{aligned} \text{ELPD} &= \sum_{i=1}^M \int_{\mathbb{R}} \log p(\bar{y}_i|\mathbf{y}) p_{\text{true}}(\bar{y}_i) d\bar{y}_i \\ &= \sum_{i=1}^M \int_{\mathbb{R}} \log \left[\int_{\Theta} \ell(\bar{y}_i|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right] p_{\text{true}}(\bar{y}_i) d\bar{y}_i. \end{aligned} \quad (192)$$

Note that $p_{\text{true}}(\bar{y}_i)$ is the density representing the true data generating process for \bar{y}_i , which is clearly unknown. Therefore, some approximations are required. First all, we define an over-estimation of the ELPD, considering the observed data in $\mathbf{y} = [y_1, \dots, y_{D_y}]$ instead new alternative data \bar{y}_i , so that $M = D_y$ and $\int_{\mathbb{R}} \log p(\bar{y}_i|\mathbf{y}) p_{\text{true}}(\bar{y}_i) d\bar{y}_i \approx \log p(y_i|\mathbf{y})$, i.e.,

$$\widehat{\text{ELPD}} = \sum_{i=1}^{D_y} \log p(y_i|\mathbf{y}) = \sum_{i=1}^{D_y} \log \left[\int_{\Theta} \ell(y_i|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right]. \quad (193)$$

In practice, we need an additional approximation for computing $p(y_i|\mathbf{y}) = \int_{\Theta} \ell(y_i|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. We can use MCMC samples from $P(\boldsymbol{\theta}|\mathbf{y})$, i.e.,

$$\widehat{\text{ELPD}} = \sum_{i=1}^{D_y} \log \hat{p}(y_i|\mathbf{y}) = \sum_{i=1}^{D_y} \log \left[\frac{1}{N} \sum_{n=1}^N \ell(y_i|\boldsymbol{\theta}_n) \right], \quad \text{with } \boldsymbol{\theta}_n \sim P(\boldsymbol{\theta}|\mathbf{y}). \quad (194)$$

LOO-CV. However, we know that the approximation above overestimates ELPD. One possibility is to use cross-validation (CV), such as the leave-one-out cross-validation (LOO-CV). In LOO-CV, we consider $p(y_i|\mathbf{y}_{-i})$ instead of $p(y_i|\mathbf{y})$ in Eq. (193), where \mathbf{y}_{-i} is vector \mathbf{y} leaving out the i -th data, y_i . Hence,

$$\widehat{\text{ELPD}}_{\text{LOO-CV}} = \sum_{i=1}^{D_y} \log p(y_i|\mathbf{y}_{-i}) = \sum_{i=1}^{D_y} \log \left[\int_{\Theta} \ell(y_i|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}_{-i}) d\boldsymbol{\theta} \right]. \quad (195)$$

For approximating $p(y_i|\mathbf{y}_{-i}) = \int_{\Theta} \ell(y_i|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{y}_{-i})d\boldsymbol{\theta}$, we draw again from the full posterior by means of an MCMC technique, $\boldsymbol{\theta}_n \sim P(\boldsymbol{\theta}|\mathbf{y})$, and apply importance sampling [94],

$$p(y_i|\mathbf{y}_{-i}) \approx \widehat{p}(y_i|\mathbf{y}_{-i}) = \sum_{n=1}^N \bar{w}_{i,n} \ell(y_i|\boldsymbol{\theta}_n), \quad \boldsymbol{\theta}_n \sim P(\boldsymbol{\theta}|\mathbf{y}), \quad (196)$$

where $\bar{w}_{i,n} = \frac{w_{i,n}}{\sum_{k=1}^N w_{i,k}}$ and, in the case the data are conditionally independent,

$$w_{i,n} = \frac{1}{\ell(y_i|\boldsymbol{\theta}_n)} \propto \frac{P(\boldsymbol{\theta}_n|\mathbf{y}_{-i})}{P(\boldsymbol{\theta}_n|\mathbf{y})}.$$

Thus, replacing in (196), we obtain

$$p(y_i|\mathbf{y}_{-i}) \approx \widehat{p}(y_i|\mathbf{y}_{-i}) = \frac{1}{\sum_{n=1}^N \frac{1}{\ell(y_i|\boldsymbol{\theta}_n)}}, \quad \boldsymbol{\theta}_n \sim P(\boldsymbol{\theta}|\mathbf{y}), \quad (197)$$

which resembles the harmonic mean estimator but with just one data point. However, since the full posterior $P(\boldsymbol{\theta}_n|\mathbf{y})$ has smaller variance of $P(\boldsymbol{\theta}_n|\mathbf{y}_{-i})$, the direct use of (197) is quite unstable, since the IS weights can have high or infinite variance. See [94] for stable computations of LOO-CV and using posterior simulations. Moreover, see also [93] for a quantitative comparison of methods for estimating the predictive ability of a model. The marginal likelihood can also be interpreted as a measure of predictive performance [87, Sect. 3.2]. In [95], the authors show that the marginal likelihood is equivalent, in some sense, to a leave-p-out cross-validation procedure. For further discussions about model selection strategies, see [96, 97].

8 Numerical comparisons

In this section, we compare the performance of different marginal likelihood estimators in different experiments. First of all, we consider 3 different illustrative scenarios in Section 8.1, 8.2 and 8.3 each one considering different challenges: different overlap between prior and likelihood (changing the number of data, or the variance and mean of the prior), multi-modality and different dimensions of the inference problem. The first experiment also considers two different sub-scenarios. Additional theoretical results related to the experiments in Sect. 8.1 are provided in the Supplementary Material.

The last two experiments involves a real data analysis. In Section 8.4, we test several estimators in a nonlinear regression problem with real data (studied also in [30]), where the likelihood function has non-elliptical contours. Finally, in Section 8.5 we consider another regression problem employing non-linear localized bases with real data of the COVID-19 outbreak.

8.1 First experiment

8.1.1 First setting: Gaussians with same mean and different variances

In this example, our goal is to compare by numerical simulations different schemes for estimating the normalizing constant of a Gaussian target $\pi(\theta) = \exp(-\frac{1}{2}\theta^2)$. We know the ground-truth

$Z = \int_{-\infty}^{\infty} \pi(\theta) d\theta = \sqrt{2\pi}$, so $P(\theta) = \frac{\pi(\theta)}{Z} = \mathcal{N}(\theta|0, 1)$. Since this is a data-independent example, $\pi(\theta)$ and $P(\theta)$ have no dependence on \mathbf{y} . We compare several estimators enumerated below, considering one or two proposals.

One proposal estimators (IS and RIS). First of all, we recall that the IS vers-1 estimator with importance density $\bar{q}(\theta)$ and the RIS estimator with auxiliary density $f(\theta)$ are

$$\hat{Z}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^N \frac{\pi(z_i)}{\bar{q}(z_i)}, \quad z_i \sim \bar{q}(\theta), \quad \hat{Z}_{\text{RIS}} = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{f(\theta_i)}{\pi(\theta_i)}}, \quad \theta_i \sim P(\theta).$$

For a fair comparison, we consider

$$\bar{q}(\theta) = f(\theta) = \mathcal{N}(\theta|0, h^2) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{1}{2h^2}\theta^2\right).$$

where $h > 0$ is the standard deviation. We desire to study the performance of the two estimators as h varies. Moreover, a theoretical comparison of IS and RIS estimators is given in the Supplementary Material.

Estimators using with two proposals. The IS and RIS estimators use a single set of samples from $\bar{q}(\theta)$ or $P(\theta)$, respectively. Now, we consider the comparison, in terms of MSE, against several estimators that use sets of samples from both densities, $\bar{q}(\theta)$ and $P(\theta)$, at the same time. Let $\{z_i\}_{i=1}^M$ and $\{\theta_j\}_{j=1}^N$ denote sets of iid samples from $\bar{q}(\theta)$ and $P(\theta)$, respectively. When $M = N = 500$, the set $\{\{z_i\}_{i=1}^M, \{\theta_j\}_{j=1}^N\}$ can be considered as a unique set of samples drawn from the mixture $\frac{1}{2}P(\theta) + \frac{1}{2}\bar{q}(\theta)$ [51]. For a fair comparison, these estimators use $\frac{M}{2}$ samples from $\bar{q}(\theta)$ and $\frac{N}{2}$ samples from $P(\theta)$.

Ideal and realistic scenarios. Furthermore, we consider two scenarios, corresponding to whether we can evaluate $P(\theta)$ (ideal and impossible scenario) or we evaluate $\pi(\theta) \propto P(\theta)$ (realistic scenario). Note that the first scenario is simply for illustration purposes.

Jointly with IS and RIS estimator, we test several other estimators of Z , introduced in Section 4.2, that use two sets of samples simultaneously.

- **Opt-BS:** The optimal bridge sampling estimator with $\alpha(\theta) = (\frac{1}{2}P(\theta) + \frac{1}{2}\bar{q}(\theta))^{-1}$.
- **Mix-IS:** IS vers-1 with the mixture $\frac{1}{2}P(\theta) + \frac{1}{2}\bar{q}(\theta)$, instead of $\bar{q}(\theta)$, as proposal.
- **Mix-self IS:** The self-IS estimator, with $f(\theta) = \bar{q}(\theta)$, and the mixture $\frac{1}{2}P(\theta) + \frac{1}{2}\bar{q}(\theta)$ as the proposal.

Moreover, we consider another one proposal estimator, described in Section 4.1.3:

- **Opt-self IS:** The optimal self-IS estimator, with $f(\theta) = \bar{q}(\theta)$. Note that this estimator use samples from a density to $\bar{q}^{\text{opt}}(\theta) \propto |P(\theta) - \bar{q}(\theta)|$. We include it as a reference, for its optimality, and since $\bar{q}^{\text{opt}}(\theta)$ involves both, $P(\theta)$ and $\bar{q}(\theta)$.

Remark 24. Clearly, in the realistic scenario, all of the schemes above must be replaced by their iterative versions, since we cannot evaluate $P(\theta)$ but only $\pi(\theta) \propto P(\theta)$.

Results in ideal scenario. Figures 8(a)-(b) show the MSE of the estimators versus h (which is the standard deviation of $\bar{q}(\theta)$) in the ideal scenario. IS vers-1 can have very high MSE when $h < 1$, i.e., $\bar{q}(\theta)$ has smaller variance than the $P(\theta)$. Whereas, IS vers-1 is quite robust when $h > 1$. The MSE of RIS has the opposite behavior of IS vers-1. This is because RIS needs that $\bar{q}(\theta)$ has lighter tails than $P(\theta)$. In this example, optimal bridge sampling seems to provide performance in-between the IS and RIS estimators. The MSE of Opt-BS is closer to RIS for $h < 1$, whereas Opt-BS becomes closer to IS for $h > 1$. Conversely, the MSE of Opt BS is not smaller than that of IS or RIS for any h in this example. Finally, Mix-IS and Mix-self-IS provide the best performance, even better than the optimal self-IS estimator. But this is due to we are in an ideal, unrealistic scenario.

Results in the realistic scenario. Since Z is unknown we cannot evaluate $P(\theta)$ but only $\pi(\theta) \propto P(\theta)$. Only IS and RIS can be truly applied. The rest of above estimators must employ an iterative procedure (see Section 4.1.3 and Section 4.2). The iterative versions of these estimators evaluate $\frac{1}{2}\pi(\theta)/\hat{Z}^{(t)} + \frac{1}{2}\bar{q}(\theta)$, where $\hat{Z}^{(t)}$ is the current approximation. In Figure 9, we show these three estimators after $T = 5$ and $T = 15$ iterations. Interestingly, note that they all converge to the results of Opt-BS estimator. This means that the iterative versions of Mix-IS and Mix-self-IS are two alternative of Opt-BS in practice, and the performance obtained in ideal scenario are unachievable. However, the iterative version of Opt-BS seems to have the fastest convergence (to the results of the ideal Opt-BS), w.r.t. the iterative versions of Mix-IS and Mix-self-IS.

We also include a two-stage version of the Opt-selfIS estimator (see Section 4.1.2). This estimator employs $\frac{N}{4}$ to obtain an approximation \hat{Z} via standard IS, and then draws $\frac{3}{4}N$ samples from a density proportional to $|\pi(\theta)/\hat{Z} - \bar{q}(\theta)|$. This two-stage Opt-selfIS depends on the quality of the initial approximation of \hat{Z} . Since this initial approximation is provided by IS, and since IS is problematic when $h < 1$, the two-stage self-IS does not perform better than Opt-BS for $h < 1$.

8.1.2 Second setting: Gaussians with same variance and different means

In this setting, we consider again $P(\theta) \propto \pi(\theta) = \exp(-\frac{1}{2}\theta^2)$, i.e., $P(\theta) = \frac{\pi(\theta)}{Z} = \mathcal{N}(\theta|0, 1)$, but the proposal is $\bar{q}(\theta) = \mathcal{N}(\theta|\mu, 1)$ for $\mu \geq 0$. Namely, as μ grows, $\bar{q}(\theta)$ and $P(\theta)$ are more distant. A theoretical comparison of IS and RIS estimators is given in the Supplementary Material, also for this setting.

Similarly, we compare the MSE as function of μ of different estimators of Z : (a) IS vers-1, (b) RIS, (c) optimal BS (Opt-BS), (d) a suboptimal self-IS estimator with $f(\theta) = \bar{q}(\theta)$ and using $\bar{q}(\theta) = \mathcal{N}(\frac{\mu}{2}, 1)$ as proposal, and (e) the Opt-self IS estimator with $f(\theta) = \bar{q}(\theta)$ and proposal $\bar{q}^{\text{opt}}(\theta) \propto |P(\theta) - \bar{q}(\theta)|$. Each estimator is computed using 500 samples in total and the results are averaged over 2000 independent simulations.

Results of the second setting. Unlike in the first setting, here we consider only the ideal

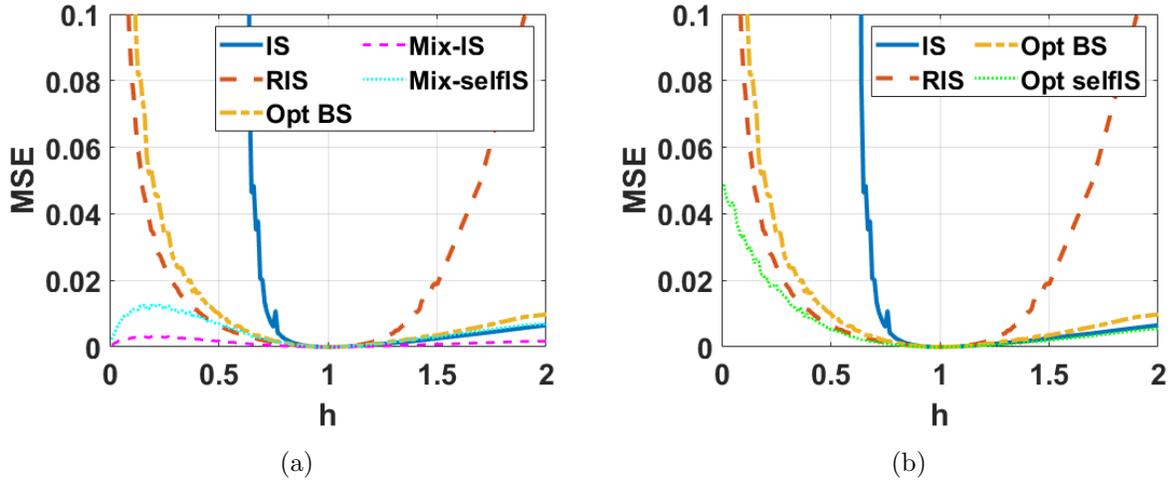


Figure 8: Numerical comparison with estimators using samples from $\bar{q}(\theta)$ and $P(\theta)$, and optimal self-IS. The figure shows the MSE of each method (averaged over 2000 simulations) as a function of h .

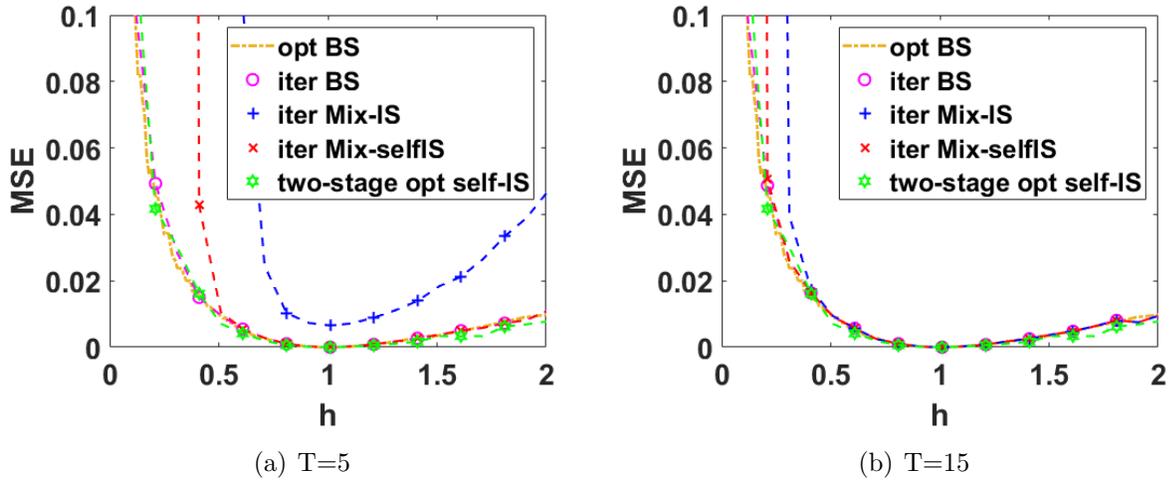


Figure 9: Comparison of iterative version of estimators with a very far starting value, $\hat{Z}^{(0)} = 5000$, with $T = 5$ and $T = 15$. Note that the two-stage self-IS is not iterative (see Section 4.1.2).

scenario (i.e., without iterative procedures). However, note that the suboptimal self-IS scheme would not require an iterative version. The results are shown in Figure 10. The MSE of both IS and RIS diverge as e^{μ^2} . Opt-BS shows better performance than IS vers-1 and RIS. The suboptimal self-IS estimator performs similarly to the Opt-BS, but both are worse than the Opt-self IS estimator. In this example, the estimators that use a middle density (as Opt-BS and the self-IS estimators) are less affected by the problem of $P(\theta)$ and $\bar{q}(\theta)$ becoming further apart. As in the previous setting, we expect that the iterative versions of Opt-BS converges to the results of the ideal Opt-

BS, provided in Figure 10. Recall that, for approximating the Opt-self IS, we require a two-stage procedure. However, a procedure with just two stages could be not enough, as we showed in the previous setting. Hence, an iterative application of the two-stage procedure could be employed (becoming actually an adaptive importance sampler).

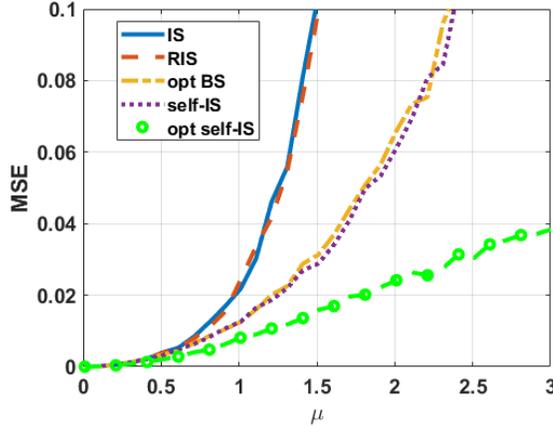


Figure 10: Numerical comparison of IS, RIS, Opt-BS, suboptimal self-IS and Opt self-IS. The figure shows the MSE of each method (averaged over 2000 simulations) as a function of μ . Greater μ means $P(\theta)$ and $\bar{q}(\theta)$ are further apart.

8.2 Second experiment: Gaussian likelihood and uniform prior

Let us consider the following one-dimensional example. More specifically, we consider independent data $\mathbf{y} = [y_1, \dots, y_{D_y}]$ generated according to a Gaussian observation model,

$$\ell(\mathbf{y}|\theta) = \prod_{i=1}^{D_y} \ell(y_i|\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^{D_y}} \exp \left\{ -\frac{D_y}{2\sigma^2} [(\theta - \bar{y}) + s_y] \right\},$$

where $\sigma = 3$, \bar{y} and s_y denote the sample mean and sample variance of \mathbf{y} , respectively. We consider a uniform prior $g(\theta) = \frac{1}{2\Delta}$, $\theta \in [-\Delta, \Delta]$ with $\Delta > 0$ being the prior width. In this setting, the marginal likelihood Z can be obtained in closed-form as a function of Δ and n (considering the evaluation of the error function $\text{erf}(x)$). The posterior is a truncated Gaussian $P(\theta|\mathbf{y}) \propto \mathcal{N}(\theta|\bar{y}, \frac{\sigma^2}{D_y})$, $\theta \in [-\Delta, \Delta]$. Let $\beta \in [0, 1]$ denote an inverse temperature, the power posterior is

$$P(\theta|\mathbf{y}, \beta) \propto \mathcal{N} \left(\theta \middle| \bar{y}, \frac{\sigma^2}{D_y \beta} \right), \quad \text{restricted to } \theta \in [-\Delta, \Delta]. \quad (198)$$

For any β , we can sample $P(\theta|\mathbf{y}, \beta) \propto \ell(\mathbf{y}|\theta)^\beta g(\theta)$ with rejection sampling by drawing from $\mathcal{N}(\theta|\bar{y}, \frac{\sigma^2}{D_y \beta})$ and discarding the samples that fall outside $[-\Delta, \Delta]$.

Scenario 1: $\Delta = 10$ and $D_y = 10$. We start by setting $\Delta = 10$ and generating $D_y = 10$ data points from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 3$. The value of the marginal likelihood is $\log Z = -25.2853$. We aim to compare the performances of several methods in estimating $\log Z$: (a) Naive Monte Carlo (NMC), (b) Harmonic mean (HM), (c) IS with a tempered posterior as proposal (IS-P), (d) stepping stone sampling (SS), (e) power posterior method (PP), and (f) path sampling (PS).

Remark 25. *Estimating $\log Z$, instead of directly Z , helps the methods of PP and PS, with respect to NMC, HM, IS-P and SS (making their results worse).*

We establish a total budget of $N = 10^3$ likelihood evaluations. For SS and PP, we set $K + 1$ values of β , from $\beta_0 = 0$ to $\beta_K = 1$, chosen (i) uniformly, i.e., $\beta_k = \frac{k}{K}$ for $k = 1, \dots, K$, or (ii) concentrated around $\beta = 0$, i.e., $\beta_k = \left(\frac{k}{K}\right)^{1/\alpha}$ with $\alpha = 0.25$. Hence the uniform case is obtained when $\alpha = 1$. Note that SS draws samples from K distributions, while PP draw samples from $K + 1$ distributions. For fair comparison, we sample $\lfloor \frac{N}{K} \rfloor$ times from each $P(\theta|\mathbf{y}, \beta_k)$, for $k = 0, \dots, K - 1$, in SS, and $\lfloor \frac{N}{K+1} \rfloor$ times from of each $P(\theta|\mathbf{y}, \beta_k)$, for $k = 0, \dots, K$, in PP. For IS-P we test $\beta_1 = 0.5$ and $\beta_2 = 0.5^4$ and draw N samples from each of the $P(\theta|\mathbf{y}, \beta_1)$ and $P(\theta|\mathbf{y}, \beta_2)$. For PS, we sample N pairs (β', θ') as follows: we first sample β' from a $\mathcal{U}(0, 1)$ and then sample θ' from the corresponding power posterior $P(\theta|\mathbf{y}, \beta')$. Naive Monte Carlo uses N independent samples from prior and HM uses N independent samples from the posterior.

Results scenario 1. In Figure 11(a), we show 500 independent estimations from each method. We observe that NMC works very well in this scenario since the prior acts as a good proposal. SS with $K = 2$ provides also good performance, since half the samples come from the prior with this choice of K . The value of α seems to be not important for SS in this case. PS performs as well as NMC and SS, but shows a slightly bigger dispersion. HM tends to overestimate the marginal likelihood, which is a well-known issue. The estimation provided by IS-P depends on the choice of β . For $\beta_1 = 0.5$, the power posterior is closer to the posterior so its behavior is similar to HM. For $\beta_2 = 0.0625$ the power posterior is close to the prior, and IS-P tends to underestimate Z . Recall that IS-P has a bias since it is a special case of IS vers-2. PP performs poorly with $K = 2$, due to the discretization error in (103), which improves when considering the value $K = 35$. The choice $\alpha = 0.25$, w.r.t. $\alpha = 1$, improves the performance in PP.

In Figure 11(b), we show the mean absolute error (MAE) in estimating $\log Z$ of SS and PP as a function of K . We depict two curves for each method, corresponding to the choices $\alpha = 1$ and $\alpha = 0.25$. We can observe that the errors obtained in SS and PP when $\alpha = 0.25$ are smaller than when $\alpha = 1$ for any K . This is in line with the recommendations provided in their original works. We note that the error of SS slightly deteriorates as K grows: for $K > 2$, less and less samples are drawn from the prior, which is a good proposal in this scenario (with $\Delta = 10$ and $D_y = 10$). The performance of PP improves drastically as K grows, since larger K means that the trapezoidal rule is more accurate in approximating (103). SS and PP, for $\alpha = 1$ and $\alpha = 0.25$, approach the same limit when K grows, achieving an error which is always greater than the one obtained by NMC, in this scenario.

Scenario 2: $\Delta = 1000$ and $D_y = 100$. Now, we replicate the previous experiment increasing the

number of data, $D_y = 100$, and the width of the prior, $\Delta = 1000$. The joint effect of increasing D_y and Δ makes the likelihood become extremely concentrated w.r.t. the prior, hence decreasing the value of the marginal likelihood, being $\log Z = -267.6471$. Moreover, this high discrepancy between prior and posterior is reflected in the power posteriors $P(\theta|\mathbf{y}, \beta)$, which will be very similar to the posterior except for very small values of β . We compare all the methods described before with a total budget of $N = 10^3$ likelihood evaluations. Additionally, we also test a PS where $\beta' \sim \mathcal{B}(0.25, 1)$, i.e., from a beta distribution which provides more β' values closer to 0.

Results scenario 2. In Figure 12(a), we can see that, unlike in the previous scenario, the NMC tends to underestimate the marginal likelihood, since the likelihood is much more concentrated than the prior. The HM and the two implementations of IS-P provide similar results, overestimating Z : in this case, the posterior is so different from the prior that $P(\theta|\mathbf{y}, 0.5)$ and $P(\theta|\mathbf{y}, 0.06)$ are very similar to the posterior. PS with $\beta' \sim [0, 1]$ tends to overestimate Z : since the β' 's are drawn uniformly in $[0, 1]$, many samples (β', θ') are drawn in high-valued likelihood zones. Indeed, at least the bias is reduced when we test PS with $\beta' \sim \mathcal{B}(0.25, 1)$. We also show the results of one implementation of SS (with $K = 10$ and $\alpha = 0.25$) and PP (with $K = 70$ and $\alpha = 0.25$). Both greatly outperform the rest of estimators in this scenario, providing accurate estimations. In Figure 12(b), we show again the MAE of SS and PP as a function of K for two values $\alpha = 1$ and $\alpha = 0.25$. The error of PP, with either $\alpha = 1$ or $\alpha = 0.25$, decreases as K grows, although it decreases more rapidly when considering $\alpha = 0.25$. The error of SS with $\alpha = 0.25$ decreases as K grows, but increases with K when $\alpha = 1$. Again, PP requires the use bigger values of K with respect to SS. In both methods, the choice of $\alpha < 1$, i.e., concentrating β 's near $\beta = 0$ where $P(\theta|\mathbf{y}, \beta)$ is usually changing rapidly, shows to improve the overall performance.

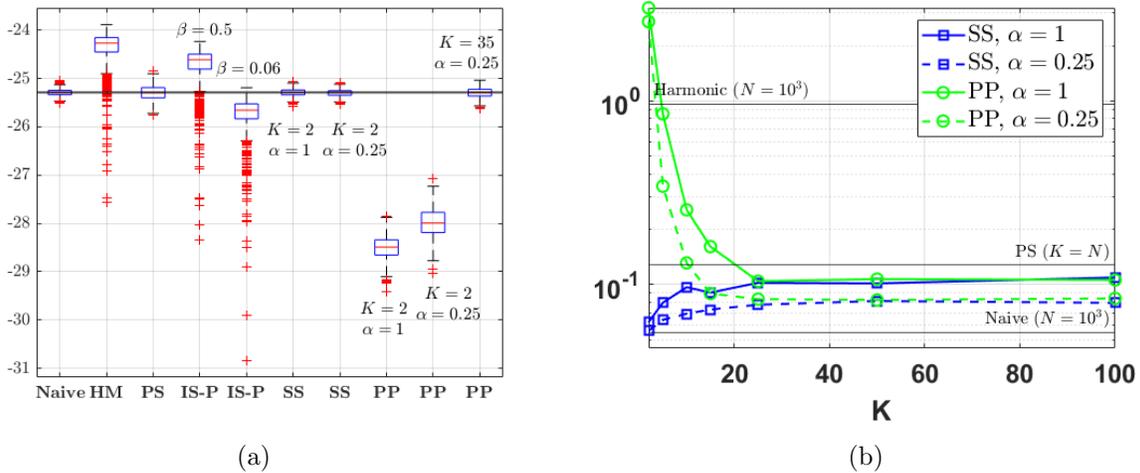


Figure 11: Simulations when $D_y = 10$ and $\Delta = 10$: (a) Estimates of $\log Z$ in 500 independent simulations, (b) MAEs of SS and PP as a function of K for two values of α .

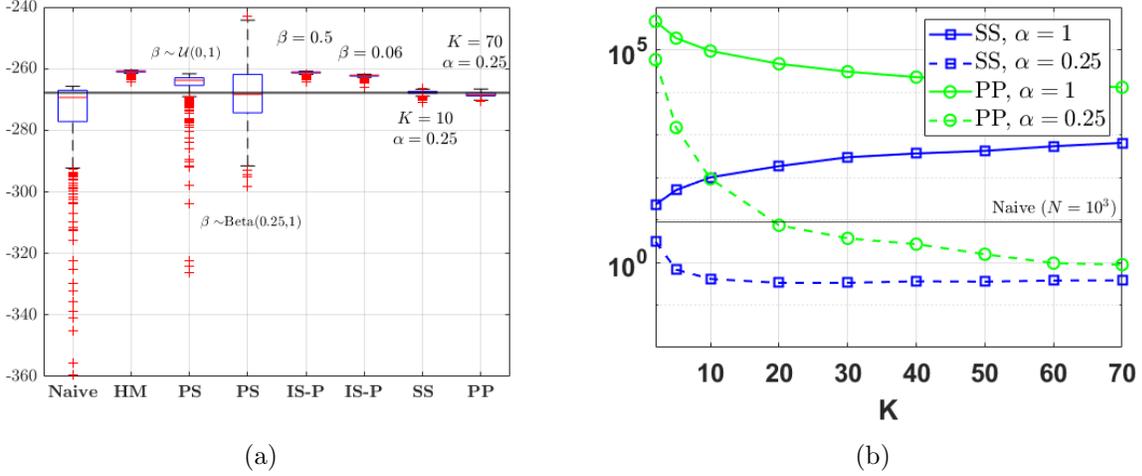


Figure 12: Simulations when $D_y = 100$ and $\Delta = 1000$: (a) Estimates of $\log Z$ in 500 independent simulations, (b) MAEs of SS and PP as a function of K for two values of α .

8.3 Third experiment: posterior as mixture of two components

We consider a posterior which is a mixture of two D_θ -dimensional Gaussian densities. It is a conjugate model where the likelihood is Gaussian and the prior is a mixture of two Gaussian. Given the observation vector \mathbf{y} , we consider a D_θ -dimensional Gaussian likelihood function

$$\ell(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\Lambda}), \quad (199)$$

with covariance $\boldsymbol{\Lambda}$, and a D_θ -dimensional Gaussian mixture prior

$$g(\boldsymbol{\theta}) = \alpha_{\text{prior}} \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\text{pr}}^{(1)}, \boldsymbol{\Sigma}_{\text{pr}}^{(1)}) + (1 - \alpha_{\text{prior}}) \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\text{pr}}^{(2)}, \boldsymbol{\Sigma}_{\text{pr}}^{(2)}), \quad (200)$$

with $\alpha_{\text{prior}} \in [0, 1]$, $\boldsymbol{\mu}_{\text{pr}}^{(i)}$ and $\boldsymbol{\Sigma}_{\text{pr}}^{(i)}$ being the prior means and covariances of each component of the mixture, respectively. Then, the posterior is also a mixture of two Gaussian densities

$$P(\boldsymbol{\theta}|\mathbf{y}) = \alpha_{\text{post}} \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\text{post}}^{(1)}, \boldsymbol{\Sigma}_{\text{post}}^{(1)}) + (1 - \alpha_{\text{post}}) \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\text{post}}^{(2)}, \boldsymbol{\Sigma}_{\text{post}}^{(2)}), \quad (201)$$

where the parameters $\alpha_{\text{post}} \in [0, 1]$, $\boldsymbol{\mu}_{\text{post}}^{(i)}$ and $\boldsymbol{\Sigma}_{\text{post}}^{(i)}$ can be obtained in closed-form from α_{prior} , $\boldsymbol{\mu}_{\text{pr}}^{(i)}$, $\boldsymbol{\Sigma}_{\text{pr}}^{(i)}$, $\boldsymbol{\Lambda}$ and \mathbf{y} . Thus, having the analytical expression of the posterior in closed-form allows to compute exactly the marginal likelihood Z (recall $Z = \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{P(\boldsymbol{\theta}|\mathbf{y})}$ for any $\boldsymbol{\theta}$). In this case, we can also draw samples directly from the posterior. We can interpret this scenario as the use of an ideal MCMC scenario, where the performance is extremely good. We compare different estimators of Z changing the Euclidean distance between the means of posterior mixture components,

$$\text{dist} = \|\boldsymbol{\mu}_{\text{post}}^{(1)} - \boldsymbol{\mu}_{\text{post}}^{(2)}\|_2, \quad (202)$$

in $D_\theta = 1$ and $D_\theta = 5$. This distance can be controlled by changing the distance between the prior modes. More specifically, we choose $\boldsymbol{\Lambda} = 50\mathbf{I}_D$, $\boldsymbol{\Sigma}_{\text{pr}}^{(1)} = \boldsymbol{\Sigma}_{\text{pr}}^{(2)} = 30\mathbf{I}_D$, where \mathbf{I}_D

denotes the D -dimensional identity matrix. The data is a single observation $\mathbf{y} = -0.5\mathbf{1}_D$, where $\mathbf{1}_D$ a D -dimensional vector of 1's. For the prior means we chose $\boldsymbol{\mu}_{\text{pr}}^{(1)} = -\boldsymbol{\mu}_{\text{pr}}^{(2)} = L\mathbf{1}_D$, so $\|\boldsymbol{\mu}_{\text{pr}}^{(1)} - \boldsymbol{\mu}_{\text{pr}}^{(2)}\|_2 = 2L\sqrt{D_\theta}$. We can change the distance between the modes of the prior, and hence between the modes of the posterior, by varying $L \in \mathbb{R}^+$. Specifically, we select $L \in \{1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51\}$ and compare: (i) the Naive-MC estimator, (ii) the HM estimator, (iii) Laplace-Metropolis estimator, (iv) RIS, and (v) CLAIS. The budget is 10^4 posterior evaluations. In RIS, we set $f(\boldsymbol{\theta})$ to be the mixture in Eq. (146), that results after applying a clustering algorithm (e.g., k-means algorithm) to the 10^4 posterior samples. In CLAIS, we use an analogous mixture obtained from $5 \cdot 10^3$ posterior samples, and then use it to draw other $5 \cdot 10^3$ samples in the lower layer (hence the total number of posterior evaluations is 10^4). For RIS and CLAIS, we set the number of clusters to $C = 4$. RIS and CLAIS also need setting the bandwidth parameter h (see Eq. (146)). We find that the choices $h = 2$ for RIS and $h = 10$ for CLAIS show the average performance of both. We test the techniques in dimension $D_\theta = 1$ and $D_\theta = 5$. We compute the relative Mean Absolute Error (MAE) in the estimation of Z , averaged over 200 independent simulations.

The results are depicted in Figure 13. They show that RIS and the CLAIS achieve the best overall performances. Their relative error remain small and rather constant for all distances considered, for $D_\theta = 1$ and $D_\theta = 5$. The RIS estimator performs as well as CLAIS in both $D_\theta = 1$ and $D_\theta = 5$, and even better for small distances in $D_\theta = 1$. For the smallest distance, the lowest relative error corresponds to the Naive MC estimator, since prior and posterior are very similar in that case, although it rapidly gets outperformed by RIS and CLAIS. The Laplace estimator provides poor results as dist grows, since the posterior becomes bimodal. As one could expect, the estimators that make use of the posterior sample to adapt its importance density, i.e., RIS and CLAIS, achieve best performances, being almost independent to increasing the distance between the modes. The HM estimator confirms its reputation of relative bad estimator.

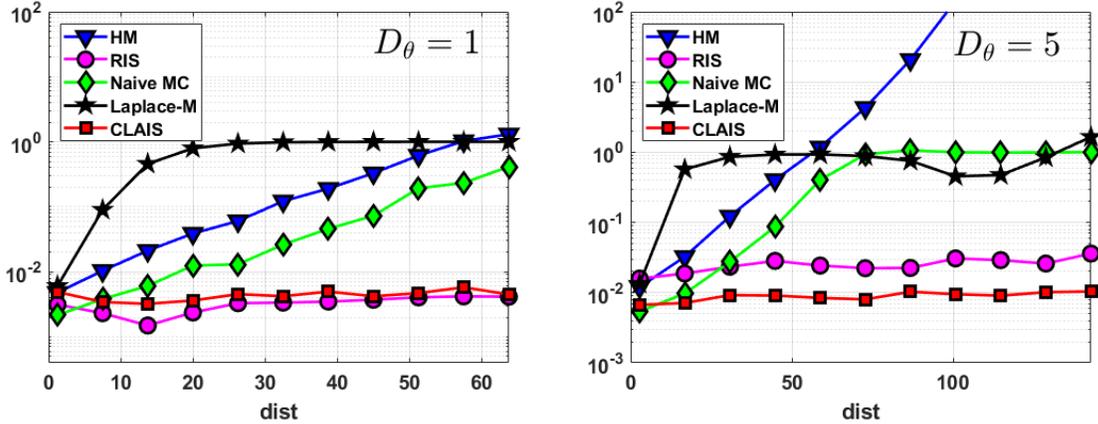


Figure 13: Relative MAE versus dist in dimension $D_\theta = 1$ and dimension $D_\theta = 5$.

8.4 Experiment with biochemical oxygen demand data

We consider a numerical experiment studied also in [30], that is a nonlinear regression problem modeling data on the biochemical oxygen demand (BOD) in terms of time instants. The outcome variable $Y_i = \text{BOD}$ (mg/L) is modeled in terms of $t_i = \text{time}$ (days) as

$$Y_i = \theta_1(1 - e^{-\theta_2 t_i}) + \epsilon_i, \quad i = 1, \dots, 6, \quad (203)$$

where the ϵ_i 's are independent $\mathcal{N}(0, \sigma^2)$ errors, hence $Y_i \sim \mathcal{N}(\theta_1(1 - e^{-\theta_2 t_i}), \sigma^2)$. The data $\{y_i\}_{i=1}^6$, measured at locations $\{t_i\}_{i=1}^6$, are shown in Table 18 below.

Table 18: Data of the numerical experiment in Section 8.4.

t_i (days)	y_i (mg/L)
1	8.3
2	10.3
3	19.0
4	16.0
5	15.6
7	19.8

The goal is to compute the normalizing constant of the posterior of $\boldsymbol{\theta} = [\theta_1, \theta_2]$ given the data $\mathbf{y} = \{(t_i, y_i)\}_{i=1}^6$. Following [30], we consider uniform priors for $\theta_1 \sim \mathcal{U}([0, 60])$, and $\theta_2 \sim \mathcal{U}([0, 6])$, i.e., $g_1(\theta_1) = \frac{1}{60}$ for $\theta_1 \in [0, 60]$, and $g_2(\theta_2) = \frac{1}{6}$, with $\theta_2 \in [0, 6]$. Moreover, we consider an improper prior for σ , $g_3(\sigma) \propto \frac{1}{\sigma}$. However, we will integrate out the variable σ . Indeed, the two-dimensional target $\pi(\boldsymbol{\theta}|\mathbf{y}) = \pi(\theta_1, \theta_2|\mathbf{y})$ results after integrating out σ by marginalizing

$$\pi(\theta_1, \theta_2, \sigma|\mathbf{y}) = \ell(\mathbf{y}|\theta_1, \theta_2, \sigma)g_1(\theta_1)g_2(\theta_2)g_3(\sigma),$$

w.r.t. σ , namely we obtain

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \int \pi(\theta_1, \theta_2, \sigma|\mathbf{y})d\sigma = \ell(\mathbf{y}|\theta_1, \theta_2)g_1(\theta_1)g_2(\theta_2) \quad (204)$$

$$= \frac{1}{60} \frac{1}{6} \frac{1}{\pi^3} \frac{8}{\left\{ \sum_{i=1}^6 [y_i - \theta_1(1 - \exp(-\theta_2 t_i))]^2 \right\}^3}, \quad [\theta_1, \theta_2] \in [0, 60] \times [0, 6], \quad (205)$$

for which we want to compute its normalizing constant $Z = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$. The derivation is given in the Supplementary Material. The true value (ground-truth) is $\log Z = -16.208$, considering the data in Table 18.

Scenario 1. As in [30], we compare the relative MAE, $\frac{\mathbb{E}[|\hat{Z}-Z|]}{Z}$, obtained by different methods: **(a)** the naive Monte Carlo estimator; **(b)** a modified version of the Laplace method (more sophisticated) given in [30]; **(c)** the Laplace-Metropolis estimator in Sect. 3.1 (using sample mean and sample covariance considering MCMC samples from $P(\boldsymbol{\theta}|\mathbf{y})$); **(d)** the HM estimator

of Eq. 46; **(f)** the RIS estimator where $f(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the MCMC samples from $P(\boldsymbol{\theta}|\mathbf{y})$ (it is denoted as RIS in Table 19); **(g)** another RIS scheme where $f(\boldsymbol{\theta})$ is obtained by a clustered KDE with $C = 4$ clusters and $h = 0$ (in a similar fashion of Eq. (146)); and, finally, a CLAIS scheme with $C \in \{1, 2\}$, $h = 0$ i.e., as in Eq. (146).

Table 19: Relative MAE, and its corresponding standard error, in estimating the marginal likelihood by seven methods

Methods	Naive	Laplace (soph)	Laplace	HM	RIS	RIS-kde	CLAIS	CLAIS
RE	0.057	0.181	0.553	0.823	0.265	0.140	0.084	0.082
std err	0.001	0.013	0.003	0.018	0.006	0.004	0.015	0.014
comments	—	see [30]	—	—	—	$C = 4$	$C = 1$	$C = 2$

All estimators consider 10000 posterior evaluations. To obtain the samples from the posterior, we run $T = 10000$ iterations of a Metropolis-Hastings algorithm, using the prior as an independent proposal pdf. The IS estimator employs 5000 posterior samples to build the normal approximation to the posterior, from which it draws 5000 additional samples. Similarly, since CLAIS draws additional samples from $\bar{q}(\boldsymbol{\theta})$ in the lower layer, in order to provide a fair comparison, we consider $N = 1$ (i.e. one chain), with $T' = T/2 = 5000$ iterations and sample 5000 additional samples in the lower layer. We averaged the relative MAE over 1000 independent runs. Our results are shown in Table 19.

In this example, and with these priors, the results show that the best performing estimator in this case is the Naive Monte Carlo, since prior and likelihood has an ample overlapping region of probability mass. However, the naive Monte Carlo scheme is generally inefficient when there is a small overlap between likelihood and prior. Note also that IS and CLAIS provide good performance. RIS-kde performs better than RIS since the choice of $f(\boldsymbol{\theta})$ in the former is probably narrower than in RIS. The worst performance is provided by the HM estimator.

Scenario 2. Now, we consider the following estimators: **(a)** the Chib’s estimator in Eq. (30), **(b)** RIS with $f(\boldsymbol{\theta})$ equal to the clustered KDE in (146) (called RIS-kde in the previous scenario), and **(c)** CLAIS with clustered KDE in (146). We study the effect of the choice of C and h in their performance. We test different numbers of clusters $C \in \{1, 2, 4, 10\}$ and different values of $h = \{0, 1, 2, 3, 4, 5\}$.

As above, we consider a fair application of CLAIS (using the same budget of posterior evaluations as in the other schemes). Moreover, in Chib’s we need to choose the point $\boldsymbol{\theta}^*$. We considered two scenarios: (i) using $\boldsymbol{\theta}^* = [19, 1]$ that is intentionally located very close to the posterior mode; (ii) using random $\boldsymbol{\theta}^*$ drawn from the priors. The first scenario clearly yields more accurate results than the second one, which we refer as a “fair” scenario (since, generally, we do not have information about the posterior modes). In summary, we compute the relative MAE of $\widehat{Z}_{\text{chib}}$, $\widehat{Z}_{\text{chib-f}}$ (where the “f” stands for “fair”), \widehat{Z}_{RIS} and $\widehat{Z}_{\text{CLAIS}}$. We compute the relative median absolute error of 1000 independent runs. Figure 14 shows the results of the experiment. CLAIS and RIS provide results, for all C and h , similar to both Chib and Chib-f. As expected, the error of $\widehat{Z}_{\text{chib}}$ is lower than

$\widehat{Z}_{\text{chib-f}}$. In CLAIS, we note that, for $C = 10$, we should not take h too small to avoid the proposal becoming problematic (i.e., narrower than the posterior). Generally, as C increases, h should not be too small since the proposal may not have fatter tails than $P(\boldsymbol{\theta}|\mathbf{y})$. The performance of RIS is best when $h = 0$, and gets worse as h increases, as expected, since $f(\boldsymbol{\theta})$ may become wider than the posterior. We expect that the results of RIS with $h = 0$ would improve further as C increases since the C-KDE pdf, in Eq. (146), will have lighter tails than the posterior. The Chib’s estimator provides also robust and good results. Overall, for the choices of C and h considered, CLAIS and RIS (with $f(\boldsymbol{\theta})$ being the clusterized KDE) provide robust results comparable to Chib’s estimator. These results are also in line with the theoretical considerations given in the Suppl. Material regarding RIS and IS.

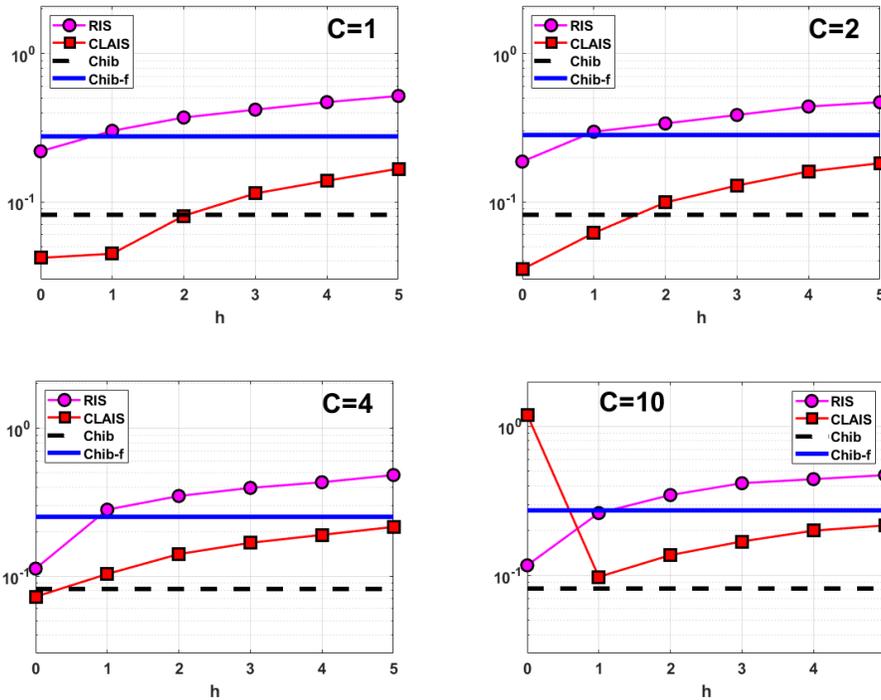


Figure 14: Relative median absolute error of RIS and CLAIS versus h for $C \in \{1, 2, 4, 10\}$. The horizontal lines correspond to Chib’s estimator (dashed) and its fair application (solid).

8.5 Experiment with COVID-19 data

Let us consider data $\mathbf{y} = [y_1, \dots, y_{D_y}]^\top$ representing the number of daily deaths caused by SAR-CoV-2 in Italy from 18 February 2020 to 6 July 2020. Let t_i denote the i -th day, we model the each observation as

$$y_i = f(t_i) + e_i, \quad i = 1, \dots, D_y = 140,$$

where f is the function that we aim to approximate and e_i 's are Gaussian perturbations. We consider the approximation of f at some t as a weighted sum of M localized basis functions,

$$f(t) = \sum_{m=1}^M \rho_m \psi(t|\mu_m, h, \nu),$$

where $\psi(t|\mu_m, h)$ is m -th basis centered at μ_m with bandwidth h . Let also be ν an index denoting the type of basis. We consider $M \in \{1, \dots, D_y\}$, then $M \leq D_y$. When $M = D_y$, the model becomes a Relevance Vector Machine (RVM), and the interpolation of all data points (maximum overfitting, with zero fitting error) is possible [98].

We consider 4 different types of basis (i.e., $\nu = 1, \dots, 4$): Gaussian ($\nu = 1$), Laplacian ($\nu = 2$), Rectangular ($\nu = 3$) and Triangular-Pyramidal ($\nu = 4$). Given ν and M , we select the locations $\{\mu_m\}_{m=1}^M$ as a uniform grid in the interval $[1, D_y]$ (recall that $D_y = 140$). Hence, knowing ν and M , the locations $\{\mu_m\}_{m=1}^M$ are given.

Likelihood and prior of $\boldsymbol{\rho}$. Let $\boldsymbol{\Psi}$ be a $D_y \times M$ matrix with elements $[\boldsymbol{\Psi}]_{i,m} = \psi(t_i|\mu_m, h)$ for $i = 1, \dots, D_y$ and $m = 1, \dots, M$, and let $\boldsymbol{\rho} = [\rho_1, \dots, \rho_M]^\top$ be the vector of coefficients, where M is the total number of bases. Then, the observation equation in vector form becomes

$$\mathbf{y} = \boldsymbol{\Psi}\boldsymbol{\rho} + \mathbf{e},$$

where \mathbf{e} is a $D_y \times 1$ vector of noise. We assume normality $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{D_y})$, where \mathbf{I}_{D_y} is the $D_y \times D_y$ identity matrix. Therefore, the likelihood function is $\ell(\mathbf{y}|\boldsymbol{\rho}, h, \sigma_e, \nu, M) = \mathcal{N}(\mathbf{y}|\boldsymbol{\Psi}\boldsymbol{\rho}, \sigma_e^2 \mathbf{I}_N)$. We also consider a Gaussian prior density over the vector of coefficients $\boldsymbol{\rho}$, i.e., $g(\boldsymbol{\rho}|\lambda) = \mathcal{N}(\boldsymbol{\rho}|\mathbf{0}, \boldsymbol{\Sigma}_\rho)$, where $\boldsymbol{\Sigma}_\rho = \lambda \mathbf{I}_M$ and $\lambda > 0$. Given ν, M, h and σ_e . Thus, the complete set of parameters is $\{\boldsymbol{\rho}, \nu, M, h, \lambda, \sigma_e\}$.

Posteriors and marginalization. With our choice of $g(\boldsymbol{\rho}|\lambda)$, the posterior of $\boldsymbol{\rho}|\lambda, h, \sigma_e$ is also Gaussian,

$$P(\boldsymbol{\rho}|\mathbf{y}, \lambda, h, \sigma_e, \nu, M) = \frac{\ell(\mathbf{y}|\boldsymbol{\rho}, h, \sigma_e, \nu, M)g(\boldsymbol{\rho}|\lambda)}{p(\mathbf{y}|\lambda, h, \sigma_e, \nu, M)} = \mathcal{N}(\boldsymbol{\rho}|\boldsymbol{\mu}_{\boldsymbol{\rho}|\mathbf{y}}, \boldsymbol{\Sigma}_{\boldsymbol{\rho}|\mathbf{y}}),$$

and a likelihood marginalized w.r.t. $\boldsymbol{\rho}$ is available in closed-form,

$$p(\mathbf{y}|\lambda, h, \sigma_e, \nu, M) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \boldsymbol{\Psi}\boldsymbol{\Sigma}_\rho\boldsymbol{\Psi}^\top + \sigma_e^2 \mathbf{I}_N). \quad (206)$$

For further details see [98]. Now, we consider priors over h, λ, σ_e , and study the following posterior

$$P(\lambda, h, \sigma_e|\mathbf{y}, \nu, M) = \frac{1}{p(\mathbf{y}|\nu, M)} p(\mathbf{y}|\lambda, h, \sigma_e, \nu, M) g_\lambda(\lambda) g_h(h) g_\sigma(\sigma_e),$$

where $g_\lambda(\lambda)$, $g_h(h)$, $g_\sigma(\sigma_e)$ are folded-Gaussian pdfs defined on $\mathbb{R}_+ = (0, \infty)$ with location and scale parameters $\{0, 100\}$, $\{0, 400\}$ and $\{1.5, 9\}$, respectively. Finally, we want to compute the marginal likelihood of this posterior, i.e.,

$$p(\mathbf{y}|\nu, M) = \int_{\mathbb{R}_+^3} p(\mathbf{y}|\lambda, h, \sigma_e, \nu, M) g_\lambda(\lambda) g_h(h) g_\sigma(\sigma_e) d\lambda dh d\sigma_e. \quad (207)$$

Furthermore, assuming a uniform probability mass $\frac{1}{D_y}$ as prior over M , we can also marginalize out M ,

$$p(M|\mathbf{y}, \nu) \propto \frac{1}{D_y} p(\mathbf{y}|\nu, M) \quad \text{and} \quad p(\mathbf{y}|\nu) = \frac{1}{D_y} \sum_{M=1}^{D_y} p(\mathbf{y}|\nu, M), \quad \text{for } \nu = 1, \dots, 4. \quad (208)$$

Considering also a uniform prior over ν , we can obtain $p(\nu|\mathbf{y}) \propto \frac{1}{4} p(\mathbf{y}|\nu)$.

For approximating $p(\mathbf{y}|\nu, M)$, for $m = 1, \dots, D_y$, we first apply a Naive Monte Carlo (NMC) method with $N = 10^4$ samples. Secondly, we run an MTM algorithm for obtaining the estimator $\widehat{Z}^{(2)}$ (see Table 13) and a Markov chain of vectors $\boldsymbol{\theta}_t = [\lambda_t, h_t, \sigma_{e,t}]$ for $t = 1, \dots, T$. This generated chain $\{\boldsymbol{\theta}_t\}_{t=1}^T$ can be also used for obtaining other estimators (e.g., the HM estimator). We consider the pairs $T = 50$, $N' = 1000$, in the MTM scheme. Therefore, $\widehat{Z}^{(2)}$ employs $N'T = 5 \cdot 10^4$ samples.

Goal. Our purpose is: (a) to make inference regarding the parameters of the model $\{\lambda, h, \sigma_e\}$, (b) approximate $Z = p(\mathbf{y}|\nu, M)$, (c) study the posterior $p(M|\mathbf{y}, \nu)$, and (d) obtain the MAP value, M_ν^* , for $\nu = 1, \dots, 4$. We also study the marginal posterior $p(\nu|\mathbf{y})$ of each of the four candidate bases.

Results. We run once NMC and MTM for all $M = 1, \dots, D_y = 140$ different models and approximate the posterior $p(M|\mathbf{y}, \nu)$ for each value of M . For illustrative reasons, in Figure 15, we show the posterior probabilities of M belonging to the intervals $[4\widetilde{M} - 3, 4\widetilde{M}]$, where \widetilde{M} is an auxiliary index $\widetilde{M} = 1, \dots, \frac{140}{4} = 35$. Thus, the first value, $\widetilde{M} = 1$ of the curves in Figure 15, represents the probability of $M \in \{1, 2, 3, 4\}$, the second value represents the probability of $M \in \{5, 6, 7, 8\}$, and so on until the last value, $\widetilde{M} = 35$, which represents the probability of $M \in \{137, 138, 139, 140\}$. We can observe that, with both techniques, we obtain that $\widetilde{M} = 2$ is the most probable interval, with a probability generally closer to 0.2, hence $M_\nu^* \in \{5, 6, 7, 8\}$. Recall that we have 35 possible intervals (values of \widetilde{M}), so when we compare with a uniform distribution $\frac{1}{35} = 0.0286$, the value 0.2 is quite high. For $\nu = 2, 3$, the corresponding probabilities are greater than 0.2, reaching 0.35 with NMC in $\nu = 2$. In Figure 16, we can observe that, with $M = 8$ bases, we are already able to obtain a very good fitting to the data.

Thus, a first conclusion is that the results obtained with models such as RVMs and Gaussian Processes (GPs) (both having $M = 140$ [98]) can be approximated in a very good way with a much more scalable model, as our model here with $M \in \{5, 6, 7, 8\}$ [98]. Regarding the marginal posterior $p(\nu|\mathbf{y})$, we can observe the results in Table 20. The basis $\nu = 3$ is discarded since is clearly not appropriate, as also shown graphically by Figure 16. With the results provided by NMC, we prefer slightly the Laplacian basis whereas, with the results of MTM, we have almost $p(\nu = 1|\mathbf{y}) \approx p(\nu = 2|\mathbf{y})$. These considerations are reasonable after having a look to Figure 16. As future work, it would be interesting to consider the locations of the bases μ_m , for $m = 1, \dots, M$, as additional parameters to be learnt.

9 Final discussion

In this work, we have provided an exhaustive review of the techniques for marginal likelihood computation with the purpose of model selection and hypothesis testing. Methods for

Table 20: The approximate marginal posterior $p(\nu|\mathbf{y})$ with different techniques.

Method	Number of used samples	$p(\nu = 1 \mathbf{y})$	$p(\nu = 2 \mathbf{y})$	$p(\nu = 3 \mathbf{y})$	$p(\nu = 4 \mathbf{y})$
NMC	10^4	0.3091	0.3307	0.0813	0.2790
MTM	$5 \cdot 10^4$	0.3155	0.3100	0.0884	0.2861

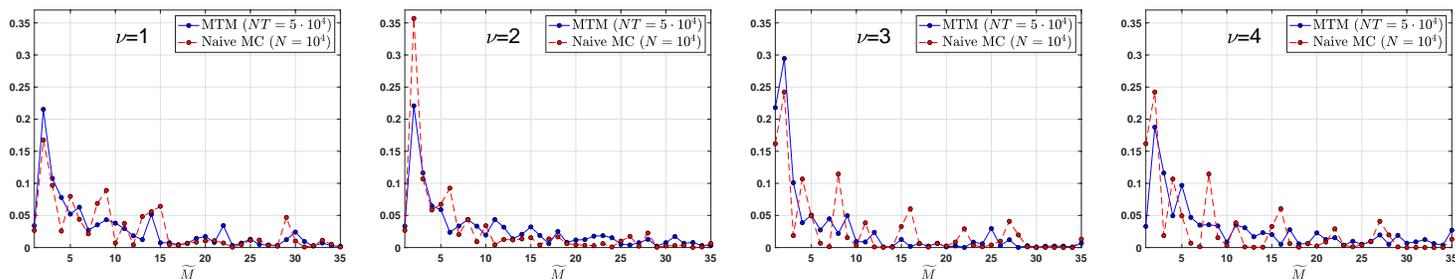


Figure 15: Posterior probabilities of the intervals $[4\widetilde{M} - 3, 4\widetilde{M}]$ with $\widetilde{M} = 1, \dots, 35$, obtained adding 4 consecutive values of $p(M|\mathbf{y}, \nu)$ with $M \in \{4\widetilde{M} - 3, 4\widetilde{M} - 2, 4\widetilde{M} - 1, 4\widetilde{M}\}$ (and $p(M|\mathbf{y}, \nu)$ is approximated by NMC or MTM). Each figure corresponds to a different type of basis, $\nu = 1, 2, 3, 4$.

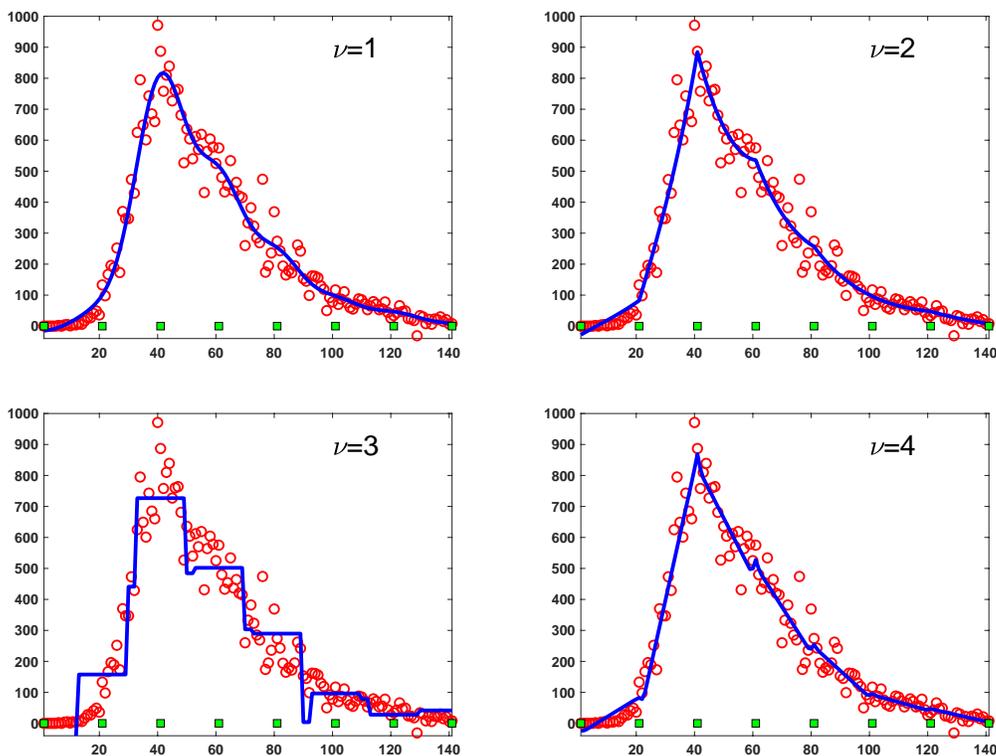


Figure 16: Best fit with 8 bases with different types of basis, $\nu = 1, 2, 3, 4$. The circles represent the analyzed data and the squares show the positions of the bases.

approximating ratios of normalizing constants have been also described. The relationships among all of them have been widely described in the text, for instance in Sections 4.2.2 and 4.3.5, by means of several summary tables (see, as examples, Tables 5, 8, and 16) and Figures from 1 to 6. The careful choice of the prior and the careful use of the improper priors in the Bayesian setting have been discussed. A brief description of alternative model selection strategies based on the posterior predictive approach, has been also provided.

Most of the presented computational techniques are based on the importance sampling (IS) approach, but also require the use of MCMC algorithms. Table 21 summarizes some methods for estimating Z , which involve the generation of the posterior $P(\boldsymbol{\theta}|\mathbf{y})$ (without using other tempered versions). This table is devoted to the interested readers which desire to obtain samples $\{\boldsymbol{\theta}_n\}_{n=1}^N$ by an MCMC method with invariant pdf $P(\boldsymbol{\theta}|\mathbf{y})$ (without either any tempering or sequence of densities) and, at the same time, also desire to approximate Z . Clearly, this table provides only a subset of all the possible techniques. They can be considered the simplest schemes, in the sense that they do not use any tempering strategy or sequence of densities. We also recall that AIC and DIC are commonly used for model comparison, although they do not directly target the actual marginal likelihood. Table 22 enumerates all the schemes that require the sampling and evaluation of tempered posteriors. For LAIS, the use of tempered posteriors is not strictly required. In PS, one could select a path that does not involve tempered posteriors. The schemes which provides unbiased estimators of Z or $\log Z$ are given in Table 23.

We also provide some final advice for practical use. First of all, if informative priors are not available, a very careful choice of the priors must be considered, as remarked in Sections 7.1 and 7.2, or alternatively a predictive posterior approach should be applied (see Section 7.4). From a computational point of view, our suggestions are listed below:

- The use of Naive Monte Carlo (NMC) should be always considered, at least as a first attempt. Moreover, the HM estimator is surely the worst estimator of Z , but it could be applied for obtaining an upper bound for Z , although it can be very imprecise/loose.
- The application of an MTM method is a good choice within the MCMC schemes. In fact, as shown in Figure 17, it also provides two estimators of Z , as well as a set of samples. These samples can be employed in other estimators, including Chib, RIS and LAIS, for instance.
- Regarding the more general task of estimating ratio of constants, In [48], the authors show that (given two unnormalized pdfs) the optimal umbrella estimator provides the best performance theoretically in estimating the ratio of their normalizing constants. However, the optimal umbrella sampling estimator is difficult and costly to implement (due to the fact sampling from the optimal umbrella proposal is not straightforward), so its best performance may not be achieved in practice.
- The Chib's method is a good alternative, that provide very good performance as we can observe in Section 8.4 and also in [99, 100]. Moreover, the Chib's method is also related to bridge sampling as discussed in Section 4.2.2. However, since it requires internal information regarding the MCMC employed (proposal, acceptance function etc.), it cannot be considered for a possible post-processing scheme after obtaining a Markov chain from a black-box MCMC algorithm. This could be easily done with the HM estimator or LAIS, for instance.

- LAIS can be considered a scheme in between the NMC and HM. NMC draw samples from the prior, which makes it rather inefficient in some setting. The HM estimator uses posterior samples but it is very unstable. LAIS uses the posterior samples to build a suitable normalized proposal, so it benefits from localizing samples in regions of high posterior probability (like the HM), while preserving the properties of standard IS (like the Naive MC). In this sense, bridge sampling, the SS method, path sampling, and the rest of techniques based on tempered posteriors, are also schemes in between the NMC and HM.
- The methods based on tempered posteriors provide very good performance but the choice of the temperature parameters β_k is important. In our opinion, among SS, PS, PP, An-IS, and SMC, the more robust to the choice of the β_k 's is the SS method (that is, perhaps, also the simplest one). Moreover, The SS method does not require the use of several tempered posteriors, unlike PS and PP. The LAIS technique can also be employed in the upper layer. Since the samples in the upper layer are only used as means of other proposal pdfs and, in the lower layer, the true posterior $P(\boldsymbol{\theta}|\mathbf{y})$ is always evaluated, LAIS is also quite robust to the choice of β_k . More comparisons among SS, An-IS, and SMC are required, since these methods are also very related as depicted in Figures 2, 4, 5 and 6.
- The nested sampling technique has gained attention and is largely applied in the literature. The derivation is complex and several approximations are considered, as discussed in Section 6.2.2. The sampling from the truncated priors is the key point and it is not straightforward [86]. In this sense, its success in the literature is surprising. However, the nested sampling includes an implicit optimization of the likelihood. We believe that is an important feature, since the knowledge of high probabilities of the likelihood is a crucial point also to the rest of computational schemes.

References

- [1] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.
- [2] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.
- [3] G. Stoltz and M. Rousset, “Free energy computations: A mathematical perspective,” *World Scientific*, 2010.
- [4] F. Liang, C. Liu, and R. Carroll, *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. England: Wiley Series in Computational Statistics, 2010.
- [5] V. Balasubramanian, “Statistical inference, Occam’s razor, and statistical mechanics on the space of probability distributions,” *Neural computation*, vol. 9, no. 2, pp. 349–368, 1997.
- [6] C. H. LaMont and P. A. Wiggins, “Correspondence between thermodynamics and inference,” *Physical Review E*, vol. 99, no. 5, p. 052140, 2019.

Table 21: Schemes for estimating Z , involving MCMC samples from $P(\boldsymbol{\theta}|\mathbf{y})$.

Method	Section	Need of drawing additional samples	Comments
Below: methods for post-processing after generating N MCMC samples from $P(\boldsymbol{\theta} \mathbf{y})$.			
Laplace	3.1	—	use MCMC for estimating $\hat{\boldsymbol{\theta}}_{\text{MAP}}$
BIC	3.2	—	use MCMC for estimating $\hat{\boldsymbol{\theta}}_{\text{MLE}}$
KDE	3.3	—	use MCMC for generating samples
Bridge	4.2.1	✓	additional samples are required; see Eq. (61)
RIS	4	—	the HM estimator is a special case
MTM	5.3	—	provides two estimators of Z
LAIS	5.4	✓	with $P(\boldsymbol{\theta} \mathbf{y})$ in the upper-layer
Below: methods that require internal information of the MCMC scheme.			
Chib’s method	3.4	✓	additional samples are required if the proposal is not independent
MTM	5.3	—	provides two estimators of Z
Below: for model selection but do not approximate the marginal likelihood			
AIC	3.2	—	use MCMC for estimating $\hat{\boldsymbol{\theta}}_{\text{MLE}}$
DIC	3.2	—	use MCMC for estimating c_p and $\bar{\boldsymbol{\theta}}$

Table 22: Methods using tempered posteriors.

Method	Section	Use of tempering strictly required
IS-P	4.3.2	without tempering, it is HM
Stepping Stones (SS)	4.3.3	✓
Path Sampling (PS)	4.3.4	other paths (without tempering) can be used
Method of Power Posteriors (PP)	4.3.6	✓
Annealed Importance Sampling (An-IS)	5.1.2	✓
Sequential Monte Carlo (SMC)	5.2	✓
Layered Adaptive Importance Sampling (LAIS)	5.4	—

- [7] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, “Akaike information criterion statistics,” *Dordrecht, The Netherlands: D. Reidel*, vol. 81, 1986.
- [8] G. Claeskens and N. L. Hjort, “The focused information criterion,” *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 900–916, 2003.
- [9] D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. der Linde, “Bayesian measures of model complexity and fit,” *J. R. Stat. Soc. B*, vol. 64, pp. 583–616, 2002.

Table 23: Methods providing unbiased estimators of Z or $\log Z$.

Method	Section
Unbiased estimators of Z :	
IS vers-1	4
Stepping Stones (SS)	4.3.3
Annealed Importance Sampling (An-IS)	5.1.2
Sequential Monte Carlo (SMC)	5.2
Layered Adaptive Importance Sampling (LAIS)	5.4
Unbiased estimators of $\log Z$:	
Path Sampling (PS)	4.3.4

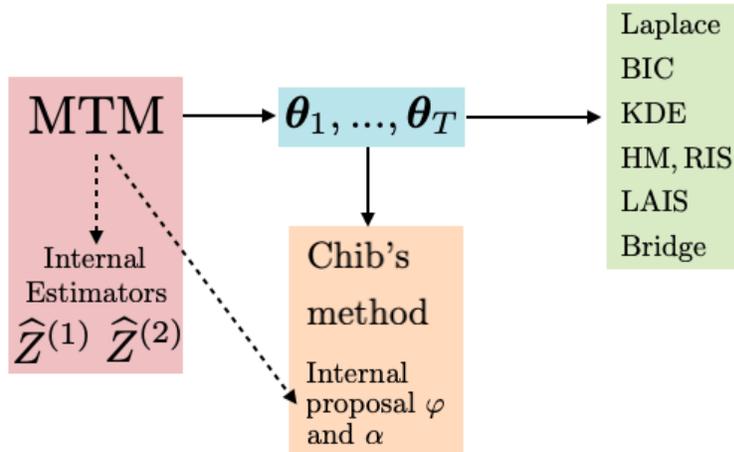


Figure 17: The application of the MTM algorithm as MCMC provides the generated samples $\{\theta_1, \dots, \theta_T\}$ and also two possible estimators of Z . The generated samples can be employed in other schemes including RIS, LAIS and Bridge sampling. Moreover, considering the proposal and the acceptance function α of the MTM, the Chib's method can be also applied. Indeed, the MTM yields a reversible chain (i.e., fulfills the balance condition).

- [10] —, “The deviance information criterion: 12 years on,” *J. R. Stat. Soc. B*, vol. 76, pp. 485–493, 2014.
- [11] C. M. Pooley and G. Marion, “Bayesian model evidence as a practical alternative to deviance information criterion,” *Royal Society Open Science*, vol. 5, no. 3, pp. 1–16, 2018.
- [12] P. Grunwald and T. Roos, “Minimum Description Length Revisited,” *arXiv:1908.08484*, pp. 1–38, 2019.
- [13] B. P. Carlin and S. Chib, “Bayesian model choice via Markov chain Monte Carlo methods,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 3, pp. 473–484, 1995.

- [14] P. J. Green, “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.
- [15] D. I. Hastie and P. J. Green, “Model choice using reversible jump Markov chain Monte Carlo,” *Statistica Neerlandica*, vol. 66, no. 3, pp. 309–338, 2012.
- [16] P. Dellaportas, J. J. Forster, and I. Ntzoufras, “On Bayesian model and variable selection using MCMC,” *Statistics and Computing*, vol. 12, no. 1, pp. 27–36, 2002.
- [17] S. J. Godsill, “On the relationship between Markov chain Monte Carlo methods for model uncertainty,” *Journal of computational and graphical statistics*, vol. 10, no. 2, pp. 230–248, 2001.
- [18] P. Congdon, “Bayesian model choice based on Monte Carlo estimates of posterior model probabilities,” *Computational statistics & data analysis*, vol. 50, no. 2, pp. 346–357, 2006.
- [19] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, “Bayesian model averaging: a tutorial,” *Statistical Science*, vol. 14, no. 4, pp. 382–417, 1999.
- [20] L. Martino, J. Read, V. Elvira, and F. Louzada, “Cooperative parallel particle filters for on-line model selection and applications to urban mobility,” *Digital Signal Processing*, vol. 60, pp. 172–185, 2017.
- [21] I. Urteaga, M. F. Bugallo, and P. M. Djurić, “Sequential Monte Carlo methods under model uncertainty,” in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, 2016, pp. 1–5.
- [22] K. H. Knuth, M. Habeck, N. K. Malakar, A. M. Mubeen, and B. Placek, “Bayesian evidence and model selection,” *Digital Signal Processing*, vol. 47, pp. 50–67, 2015.
- [23] G. Schwarz *et al.*, “Estimating the dimension of a model,” *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [24] E. J. Hannan and B. G. Quinn, “The determination of the order of an autoregression,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, 1979.
- [25] S. Chib, “Marginal likelihood from the Gibbs output,” *Journal of the american statistical association*, vol. 90, no. 432, pp. 1313–1321, 1995.
- [26] N. G. Polson and J. G. Scott, “Vertical-likelihood Monte Carlo,” *arXiv preprint arXiv:1409.3601*, 2014.
- [27] M. D. Weinberg *et al.*, “Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution,” *Bayesian Analysis*, vol. 7, no. 3, pp. 737–770, 2012.
- [28] J. Skilling, “Nested sampling for general Bayesian computation,” *Bayesian analysis*, vol. 1, no. 4, pp. 833–859, 2006.

- [29] S. M. Lewis and A. E. Raftery, “Estimating Bayes factors via posterior simulation with the LaplaceMetropolis estimator,” *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 648–655, 1997.
- [30] T. J. DiCiccio, R. E. Kass, A. Raftery, and L. Wasserman, “Computing Bayes factors by combining simulation and asymptotic approximations,” *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 903–915, 1997.
- [31] H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren, “Bayesian computing with INLA: a review,” *Annual Review of Statistics and Its Application*, vol. 4, pp. 395–421, 2017.
- [32] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.
- [33] S. Chib and I. Jeliazkov, “Marginal likelihood from the Metropolis–Hastings output,” *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 270–281, 2001.
- [34] L. Martino and V. Elvira, “Metropolis sampling,” *Wiley StatsRef: Statistics Reference Online*, pp. 1–18, 2017.
- [35] L. Martino, “A review of multiple try MCMC algorithms for signal processing,” *Digital Signal Processing*, vol. 75, pp. 134 – 152, 2018.
- [36] A. Mira and G. Nicholls, “Bridge estimation of the probability density at a point,” Department of Mathematics, The University of Auckland, New Zealand, Tech. Rep., 2003.
- [37] C. E. Rasmussen and Z. Ghahramani, “Bayesian Monte Carlo,” *Advances in neural information processing systems*, pp. 505–512, 2003.
- [38] W. R. Gilks and P. Wild, “Adaptive Rejection Sampling for Gibbs Sampling,” *Applied Statistics*, vol. 41, no. 2, pp. 337–348, 1992.
- [39] W. R. Gilks, N. G. Best, and K. K. C. Tan, “Adaptive Rejection Metropolis Sampling within Gibbs Sampling,” *Applied Statistics*, vol. 44, no. 4, pp. 455–472, 1995.
- [40] L. Martino, R. Casarin, F. Leisen, and D. Luengo, “Adaptive independent sticky MCMC algorithms,” *EURASIP Journal on Advances in Signal Processing (to paper)*, 2017.
- [41] L. Martino, J. Read, and D. Luengo, “Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling,” *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3123–3138, June 2015.
- [42] F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic, “Probabilistic integration: A role in statistical computation?” *Statistical Science*, vol. 34, no. 1, pp. 1–22, 2019.

- [43] M. H. Chen, Q. M. Shao, and J. G. Ibrahim, *Monte Carlo methods in Bayesian computation*. Springer, 2012.
- [44] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric, “Adaptive importance sampling: the past, the present, and the future,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.
- [45] A. E. Gelfand and D. K. Dey, “Bayesian model choice: asymptotics and exact calculations,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 56, no. 3, pp. 501–514, 1994.
- [46] C. P. Robert and D. Wraith, “Computational methods for Bayesian model choice,” *AIP conference proceedings*, vol. 1193, no. 1, pp. 251–262, 2009.
- [47] Y.-B. Wang, M.-H. Chen, L. Kuo, and P. O. Lewis, “A new Monte Carlo method for estimating marginal likelihoods,” *Bayesian analysis*, vol. 13, no. 2, p. 311, 2018.
- [48] M.-H. Chen, Q.-M. Shao *et al.*, “On Monte Carlo methods for estimating ratios of normalizing constants,” *The Annals of Statistics*, vol. 25, no. 4, pp. 1563–1594, 1997.
- [49] G. M. Torrie and J. P. Valleau, “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling,” *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977.
- [50] M.-H. Chen, “Importance-weighted marginal Bayesian posterior density estimation,” *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 818–824, 1994.
- [51] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, “Generalized Multiple Importance Sampling,” *Statistical Science*, vol. 34, no. 1, pp. 129–155, 2019.
- [52] X.-L. Meng and W. H. Wong, “Simulating ratios of normalizing constants via a simple identity: a theoretical exploration,” *Statistica Sinica*, pp. 831–860, 1996.
- [53] C. J. Geyer, “Estimating normalizing constants and reweighting mixtures,” *Technical Report, number 568 - School of Statistics, University of Minnesota*, 1994.
- [54] E. Cameron and A. Pettitt, “Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis,” *Statistical Science*, vol. 29, no. 3, pp. 397–419, 2014.
- [55] M. A. Newton and A. E. Raftery, “Approximate Bayesian inference with the weighted likelihood bootstrap,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 56, no. 1, pp. 3–26, 1994.
- [56] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, “Heretical multiple importance sampling,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1474–1478, 2016.
- [57] —, “Efficient multiple importance sampling estimators,” *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1757–1761, 2015.

- [58] Q. Liu, J. Peng, A. Ihler, and J. Fisher III, “Estimating the partition function by discriminance sampling,” in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015, pp. 514–522.
- [59] A. Gelman and X. L. Meng, “Simulating normalizing constants: From importance sampling to bridge sampling to path sampling,” *Statistical science*, pp. 163–185, 1998.
- [60] N. Friel and A. N. Pettitt, “Marginal likelihood estimation via power posteriors,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 3, pp. 589–607, 2008.
- [61] N. Friel, M. Hurn, and J. Wyse, “Improving power posterior estimation of statistical evidence,” *Statistics and Computing*, vol. 24, no. 5, pp. 709–723, 2014.
- [62] C. J. Oates, T. Papamarkou, and M. Girolami, “The controlled thermodynamic integral for Bayesian model evidence evaluation,” *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 634–645, 2016.
- [63] W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M. H. Chen, “Improving marginal likelihood estimation for Bayesian phylogenetic model selection,” *Systematic biology*, vol. 60, no. 2, pp. 150–160, 2010.
- [64] L. Martino, V. Elvira, and G. Camps-Valls, “Group importance sampling for particle filtering and MCMC,” *Digital Signal Processing*, vol. 82, pp. 133 – 151, 2018.
- [65] N. Chopin, “A sequential particle filter for static models,” *Biometrika*, vol. 89, pp. 539–552, 2002.
- [66] R. M. Neal, “Annealed importance sampling,” *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.
- [67] W. R. Gilks and C. Berzuini, “Following a moving target-Monte Carlo inference for dynamic Bayesian models,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 63, no. 1, pp. 127–146, 2001.
- [68] P. D. Moral, A. Doucet, and A. Jasra, “Sequential Monte Carlo samplers,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 411–436, 2006.
- [69] L. Martino, V. Elvira, and M. F. Louzada, “Effective Sample Size for importance sampling based on the discrepancy measures,” *Signal Processing*, vol. 131, pp. 386–401, 2017.
- [70] L. Martino, V. Elvira, and F. Louzada, “Weighting a resampled particle in Sequential Monte Carlo,” *IEEE Statistical Signal Processing Workshop, (SSP)*, vol. 122, pp. 1–5, 2016.
- [71] C. A. Naesseth, F. Lindsten, and T. B. Schon, “Nested Sequential Monte Carlo methods,” *Proceedings of the International Conference on Machine Learning*, vol. 37, pp. 1–10, 2015.

- [72] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert, “Population Monte Carlo,” *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 907–929, 2004.
- [73] A. Kong, “A note on importance sampling using standardized weights,” *Technical Report 348, Department of Statistics, University of Chicago*, 1992.
- [74] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez, “Particle filtering,” *IEEE Signal Processing Magazine*, vol. 20, no. 5, pp. 19–38, September 2003.
- [75] A. Doucet and A. M. Johansen, “A tutorial on particle filtering and smoothing: fifteen years later,” *technical report*, 2008.
- [76] L. Martino and J. Read, “On the flexibility of the design of multiple try Metropolis schemes,” *Computational Statistics*, vol. 28, no. 6, pp. 2797–2823, 2013.
- [77] L. Martino, V. P. D. Olmo, and J. Read, “A multi-point Metropolis scheme with generic weight functions,” *Statistics & Probability Letters*, vol. 82, no. 7, pp. 1445–1453, 2012.
- [78] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djurić, “Adaptive importance sampling: The past, the present, and the future,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.
- [79] M. F. Bugallo, L. Martino, and J. Corander, “Adaptive importance sampling in signal processing,” *Digital Signal Processing*, vol. 47, pp. 36–49, 2015.
- [80] L. Martino, V. Elvira, D. Luengo, and J. Corander, “Layered adaptive importance sampling,” *Statistics and Computing*, vol. 27, no. 3, pp. 599–623, 2017.
- [81] I. Schuster and I. Klebanov, “Markov Chain Importance Sampling—a highly efficient estimator for MCMC,” *arXiv preprint arXiv:1805.07179*, 2018.
- [82] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert, “Adaptive multiple importance sampling,” *Scandinavian Journal of Statistics*, vol. 39, no. 4, pp. 798–812, December 2012.
- [83] L. Martino and V. Elvira, “Compressed Monte Carlo for distributed Bayesian inference,” *viXra:1811.0505*, 2018.
- [84] L. Martino, V. Elvira, and D. Luengo, “Anti-tempered layered adaptive importance sampling,” *International Conference on Digital Signal Processing (DSP)*, 2017.
- [85] L. Martino, D. Luengo, and J. Míguez, “Independent random sampling methods,” *Springer*, 2018.
- [86] N. Chopin and C. P. Robert, “Properties of nested sampling,” *Biometrika*, vol. 97, no. 3, pp. 741–755, 2010.

- [87] R. E. Kass and A. E. Raftery, “Bayes factors,” *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.
- [88] D. J. Spiegelhalter and A. F. Smith, “Bayes factors for linear and log-linear models with vague prior information,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 3, pp. 377–387, 1982.
- [89] A. O’Hagan, “Fractional Bayes factors for model comparison,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 99–118, 1995.
- [90] J. O. Berger and L. R. Pericchi, “The intrinsic Bayes factor for model selection and prediction,” *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 109–122, 1996.
- [91] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. S., “A survey of monte carlo methods for parameter estimation,” *EURASIP J. Adv. Signal Process.*, vol. 25, pp. 1–62, 2020.
- [92] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.
- [93] J. Piironen and A. Vehtari, “Comparison of Bayesian predictive methods for model selection,” *Statistics and Computing*, vol. 27, no. 3, pp. 711–735, 2017.
- [94] A. Vehtari, A. Gelman, and J. Gabry, “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC,” *Statistics and computing*, vol. 27, no. 5, pp. 1413–1432, 2017.
- [95] E. Fong and C. Holmes, “On the marginal likelihood and cross-validation,” *Biometrika*, vol. 107, no. 2, pp. 489–496, 2020.
- [96] C. Alston, P. Kuhnert, L. S. Choy, R. McVinish, and K. Mengersen, “Bayesian model comparison: Review and discussion,” *International Statistical Institute, 55th session*, 2005.
- [97] R. B. O’Hara and M. J. Sillanpää, “A review of Bayesian variable selection methods: what, how and which,” *Bayesian analysis*, vol. 4, no. 1, pp. 85–117, 2009.
- [98] L. Martino and J. Read, “Joint introduction to Gaussian Processes and Relevance Vector Machines with connections to Kalman filtering and other kernel smoothers,” *arXiv:2009.09217*, pp. 1–50, 2020.
- [99] J. M. Marin and C. P. Robert, “Importance sampling methods for Bayesian discrimination between embedded models,” *arXiv preprint arXiv:0910.2325*, 2009.
- [100] N. Friel and J. Wyse, “Estimating the evidence—a review,” *Statistica Neerlandica*, vol. 66, no. 3, pp. 288–308, 2012.
- [101] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.

- [102] C. S. Bos, “A comparison of marginal likelihood computation methods,” in *Compstat*. Springer, 2002, pp. 111–116.
- [103] V. Vyshemirsky and M. A. Girolami, “Bayesian ranking of biochemical system models,” *Bioinformatics*, vol. 24, no. 6, pp. 833–839, 2007.
- [104] D. Ardia, N. Baştürk, L. Hoogerheide, and H. K. Van Dijk, “A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood,” *Computational Statistics & Data Analysis*, vol. 56, no. 11, pp. 3398–3414, 2012.
- [105] A. Schöniger, T. Wöhling, L. Samaniego, and W. Nowak, “Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence,” *Water resources research*, vol. 50, no. 12, pp. 9484–9513, 2014.
- [106] P. Liu, A. S. Elshall, M. Ye, P. Beerli, X. Zeng, D. Lu, and Y. Tao, “Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods,” *Water Resources Research*, vol. 52, no. 2, pp. 734–758, 2016.
- [107] Z. Zhao and T. A. Severini, “Integrated likelihood computation methods,” *Computational Statistics*, vol. 32, no. 1, pp. 281–313, 2017.
- [108] J. R. Oaks, K. A. Cobb, V. N. Minin, and A. D. Leaché, “Marginal likelihoods in phylogenetics: a review of methods and applications,” *Systematic biology*, vol. 68, no. 5, pp. 681–697, 2019.

Appendices

A Table of other reviews

The related literature is rather vast. In this section, we provide a brief summary that intends to be illustrative rather than exhaustive, by means of Table 24. The most relevant (in our opinion) and related surveys are compared according to the topics, material and schemes described in the work. The proportion of covering and overlapping with this work is roughly classified as “partial” \diamond , “complete” \surd , “remarkable” or “more exhaustive” work with \star . From Table 24, we can also notice the completeness of this work. We take into account also the completeness and the depth of details provided in the different derivations. The Christian Robert’s blog deserves a special mention (<https://xianblog.wordpress.com>), since Professor C. Robert has devoted several entries of his blog with very interesting comments regarding the marginal likelihood estimation and related topics.

Table 24: Covering of the considered topics of other surveys or works (\diamond : partial, \checkmark : complete, \star : remarkable or more exhaustive). We take into account also the completeness and the depth of details provided in the different derivations. To be more precise, in the case of Section 4.1, we have also considered the subsections.

Surveys	Families 1-2	IS			Advanced schemes			Vertical likelihood			Improper	
		1 prop.	2 prop.	Multiple	MCMC within IS	MTM	AIS	5.1	5.2	5.3		
Gelfand and Dey (1994)[45]	\checkmark	\diamond										
Kass and Raftery (1995)[87]		\checkmark	\diamond									
Raftery (1995)[101, Ch. 10]	\diamond	\checkmark	\diamond									
Meng and Wong (1996)[52]		\diamond	\star	\diamond								
DiCiccio et al (1997)[30]	\star	\checkmark	\checkmark									
Chen and Shao (1997)[48]	\diamond	\diamond	\checkmark	\checkmark								
Chen et al (2012)[43, Ch. 5]		\checkmark	\star	\star								
Gelman and Meng (1997)[59]	\checkmark	\checkmark										
Bos (2002)[102]	\diamond	\checkmark		\diamond	\diamond							
Vyshemirsky and Girolami (2007)[103]		\checkmark	\diamond									
Marin and Robert (2009)[99]	\diamond	\checkmark	\checkmark									
Robert and Wraith (2009)[46]		\checkmark	\checkmark								\checkmark	
Friel and Wyse (2012)[100]	\diamond	\diamond		\diamond	\checkmark						\checkmark	
Ardia et al (2012)[104]	\diamond	\diamond	\checkmark									
Polson and Scott (2014)[26]		\diamond		\diamond					\checkmark	\star	\star	
Schniger et al (2014)[105]	\checkmark	\checkmark								\diamond		
Knuth et al (2015)[22]	\diamond	\diamond		\checkmark	\checkmark					\checkmark		
Liu et al (2016)[106]	\diamond	\diamond		\checkmark						\star		
Zhao and Severini (2017)[107]	\star	\checkmark	\checkmark									
Martino (2018)[35]						\diamond	\star					
Bugallo et al (2017)[78]		\diamond						\star				
Bugallo et al (2015)[79]		\diamond										
Oaks et al (2019)[108]		\diamond		\checkmark	\diamond	\diamond				\diamond		
O'Hagan (1995)[89]												
Berger and Pericchi (1996)[90]												\star