On the computation of marginal likelihood via MCMC for model selection and hypothesis testing

Fernando Llorente[†], Luca Martino[⋆], David Delgado[†], Javier López-Santiago[†]
[†] Universidad Carlos III de Madrid, Léganes, Madrid, Spain.
[⋆] Universidad Rey Juan Carlos I, Móstoles, Madrid, Spain.

***Abstract*—In the Bayesian setting, the marginal likelihood is the key quantity for model selection purposes. Several computational methods have been proposed in the literature for the computation of the marginal likelihood. In this paper, we briefly review different estimators based on MCMC simulations. We also suggest the use of a kernel density estimation procedure, based on a clustering scheme, within some of them. Numerical comparisons are also provided.**

***Index Terms*—Bayesian evidence, marginal likelihood, Markov Chain Monte Carlo (MCMC), importance sampling**

I. Introduction

Bayesian models have become very popular in signal processing over the last decades [1], [2]. The use of Bayesian statistics in many scientific fields has spread the overall interest in estimating the so-called marginal likelihood (a.k.a., Bayesian evidence) for model selection purposes [3].

Generally, its computation is analytically intractable, so the use of advanced numerical techniques is required. The problem of computing marginal likelihoods is related to the problem of computing normalizing constants which has been of great interest in computational physics [4]. More generally, there exists a vast literature of methods for estimating the marginal likelihood. The interested readers can find complete and recent reviews in [3], [5, Chapter 5], [6]. Here, we provide a classification of the techniques distinguishing four general families that we describe below.

(a) Deterministic approaches: the idea is to replace the posterior density $\bar{\pi}(\mathbf{x})$ with a surrogate function $\widehat{\pi}(\mathbf{x})$ (obtained by an analytical approximation) whose normalizing constant is available in closed form. Examples of this class are the Laplace approximation or the Bayesian information criterion (BIC) [7].

(b) Approaches exploiting the functional identity: Consider the equality $\bar{\pi}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z}$ where Z represents the marginal likelihood, and $\pi(\mathbf{x})$ is the available unnormalized posterior. This identity shows that we can compute Z if we can estimate the posterior $\bar{\pi}(\mathbf{x})$ at least in one point $\mathbf{x}^*$, obtaining $\widehat{\pi}(\mathbf{x}^*)$. Indeed, in that case, $\widehat{Z} = \frac{\pi(\mathbf{x}^*)}{\widehat{\pi}(\mathbf{x}^*)}$. Generally, the point $\mathbf{x}^*$ is often chosen in a high-probability region. The different techniques differ by the procedure employed for obtaining the estimation $\widehat{\pi}(\mathbf{x}^*)$. In this work, we will show two alternatives belonging to this family.

(c) Approaches based on Importance Sampling (IS): These methods rely on the representation of the marginal likelihood Z as an expected value [8]. All the Monte Carlo estimators for computing Z can be considered as based on IS or its extensions [9].

(d) Approaches based on vertical representation: These methods rely on the representation of Z as one-dimensional integral and applying a quadrature procedure. Examples of this families are the nested sampling algorithm [10], [11] and vertical-likelihood Monte Carlo [12].

In this work, we focus on methods belonging to the families (b) and (c) described above, where also samples generated by an Markov Chain Monte Carlo (MCMC) scheme are employed. MCMC algorithms are widely applied in Bayesian statistics for approximating posterior distributions [8]. The success and diffusion of MCMC methods are mainly due to their intrinsic exploratory behavior. Namely, standard MCMC schemes (for instance, using a random-walk proposal density) can be easily applied to explore the state-space. However, it is well-known that MCMC methods do not provide a straightforward solution to estimate the marginal likelihood, unlike the IS schemes. Therefore, we describe and compare different marginal likelihood estimators that combine MCMC and IS algorithms in different ways. All of them are conceptually easy and straightforward to implement. We also propose and test the combination of these estimators with kernel density estimation (KDE) procedures based on clustering algorithms.

II. Problem statement and background

Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$ be the variable of interest and consider a set of observed measurements, $\mathbf{y} \subset \mathbb{R}^{D_Y}$. In the Bayesian framework, one complete model $\mathcal{M}$ is formed by a likelihood function $\ell(\mathbf{y}|\mathbf{x}, \mathcal{M})$ and an a-priori probability density function (pdf) $g(\mathbf{x}|\mathcal{M})$. All the statistical information is summarized by the posterior pdf,

$$p(\mathbf{x}|\mathbf{y}, \mathcal{M}) = \frac{\ell(\mathbf{y}|\mathbf{x}, \mathcal{M})g(\mathbf{x}|\mathcal{M})}{\int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x}, \mathcal{M})g(\mathbf{x}|\mathcal{M})d\mathbf{x}}. \tag{1}$$

The denominator plays the role of a normalizing constant and is useful for model comparison,

$$p(\mathbf{y}|\mathcal{M}) = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x}, \mathcal{M})g(\mathbf{x}|\mathcal{M})d\mathbf{x}. \tag{2}$$

The quantity above is called marginal likelihood, a.k.a., Bayesian evidence. For simplicity, we drop the dependency on the model indicator $\mathcal{M}$ hereafter and define $\bar{\pi}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \mathcal{M})$, $\pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x}, \mathcal{M})g(\mathbf{x}|\mathcal{M})$, $Z = p(\mathbf{y}|\mathcal{M})$, so that we can write

$$\bar{\pi}(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x}). \tag{3}$$

Generally, we can only evaluate $\pi(\mathbf{x})$ due to Z being unknown. Monte Carlo methods, such as IS and MCMC algorithms, allow to approximate $\bar{\pi}(\mathbf{x})$ by a cloud of samples, hence virtually any integral involving the posterior can be approximated by Monte Carlo quadrature. The IS approach also provides a simple estimator of Z. However, its efficiency depends strictly on the choice of an auxiliary proposal density. Although the MCMC methods are widely applied in the literature (mainly for their intrinsic explorative behavior), the approximation Z is not so straightforward by using MCMC samples. In this work, we describe and compare different estimators of $Z = p(\mathbf{y}|\mathcal{M})$ based on MCMC samples. Some of them are very robust and efficient and are obtained as combination of IS and MCMC schemes. Below, we introduce the two main ingredients needed to build those estimators: the Metropolis-Hastings (MH) algorithm and the kernel density estimation procedure.

A. Metropolis-Hastings (MH) algorithm

As example of MCMC method, we present the MH algorithm [8], [13]. Let $\varphi(\mathbf{x}'|\mathbf{x})$ denote a conditional density from which the proposed state $\mathbf{x}'$ is drawn conditional on the current state $\mathbf{x}$. The MH algorithm starts from an initial state $\mathbf{x}_0$ and proceeds as follows: at step t,

1) Sample $\mathbf{x}'$ from $\varphi(\mathbf{x}'|\mathbf{x}_{t-1})$.
2) Accept $\mathbf{x}'$, i.e., $\mathbf{x}_t = \mathbf{x}'$, with probability

$$\alpha(\mathbf{x}_{t-1}, \mathbf{x}') = \min\left\{1, \frac{\pi(\mathbf{x}')\varphi(\mathbf{x}_{t-1}|\mathbf{x}')}{\pi(\mathbf{x}_{t-1})\varphi(\mathbf{x}'|\mathbf{x}_{t-1})}\right\}. \quad (4)$$

Otherwise reject $\mathbf{x}'$, and set $\mathbf{x}_t = \mathbf{x}_{t-1}$.

The algorithm above yields a Markov chain with $\bar{\pi}(\mathbf{x})$ as stationary density. The resulting samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are, then, correlated but after a burn-in period will be distributed as $\bar{\pi}(\mathbf{x})$. The use of adaptive and/or advanced MCMC techniques can decrease substantially the correlation among the samples (e.g., for instance [14, Chapter 7], [13], [15], [16]).

B. Standard and Clustered KDE

The standard kernel density estimation (KDE) is a non-parametric technique to estimate a density given a set of samples. In our case, consider N independent samples, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, from $\bar{\pi}(\mathbf{x})$ [17]. Let $K_h(\mathbf{x} - \mathbf{x}_i)$ denote a kernel function with bandwidth h centered at $\mathbf{x}_i$. The KDE of $\bar{\pi}(\mathbf{x})$ is given by

$$\widehat{\pi}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} K_h(\mathbf{x} - \mathbf{x}_i). \quad (5)$$

For simplicity we consider the use of Gaussian kernels, i.e., $K_h(\mathbf{x} - \mathbf{x}_i) = \mathcal{N}(\mathbf{x}|\mathbf{x}_i, h\mathbf{I}_D)$. The optimal selection of h is a key problem. Several suggestions are provided in the literature, depending on the number of samples N, the dimension of the space D and the empirical variance obtained by the samples $\widehat{\sigma}$, such as Scott's rule $h = \widehat{\sigma}N^{-\frac{1}{D+4}}$ (when $D > 1$, then $\widehat{\sigma}$ denotes the average of the variances of each component of $\mathbf{x}$) [17, Chapter 3]. For the selection of the bandwidth h, we should also consider the correlation among the samples. However, in this work, we assume the correlation is negligible due to the use of an efficient MCMC technique [14, Chapter 7], [13]. Note that, when N is large, the evaluation of the mixture in Eq. (5) can be very costly. The problem grows if the $\widehat{\pi}(\mathbf{x})$ must be evaluated several times. For this reason and to reduce the dependence of the choice of h we consider also a *clusterized* KDE (C-KDE) procedure, formed by the following steps:

(a) Apply a clustering algorithm (e.g., k-means) or simply a compression scheme (see, e.g. [18]) to $\{\mathbf{x}_n\}_{n=1}^N$, with given number of clusters or disjoint sub-regions C.
(b) Obtain the centroids $\mathbf{m}_i$ and the covariance matrices $\mathbf{\Sigma}_i$, $i = 1, \ldots, C$ by computing the sample mean and defining $\mathbf{S}_i = \mathbf{\Sigma}_i + h\mathbf{I}_D$, respectively, where $h > 0$ and $\mathbf{I}_D$ is the $D \times D$ identity matrix.
(c) Compute the weights $\bar{w}_i = \frac{n_i}{N}$, where n_i is the number of samples belonging to i-th cluster;
(d) Hence, the final C-KDE estimation is given by

$$\widehat{\pi}(\mathbf{x}) = \sum_{i=1}^{C} \bar{w}_i \mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{S}_i). \quad (6)$$

Note that, if $C = N$, the clusters contain only one sample and Eq. (6) is equivalent to the standard KDE in Eq. (5) with Gaussian kernels.

III. Marginal likelihood estimators

In this section, we briefly describe different schemes for estimating the Bayesian evidence, which employ MCMC samples in different ways. Moreover, in most of them, we introduce the use of the clusterized KDE. We consider two families of estimators, denoted with (b) and (c) in the list of classes provided in the introduction.

A. Methods exploiting the functional equality

The identity $\bar{\pi}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z}$ shows that we can approximate Z as $\widehat{Z} = \dfrac{\pi(\mathbf{x}^*)}{\widehat{\pi}(\mathbf{x}^*)}$, where $\widehat{\pi}(\mathbf{x}^*) \approx \bar{\pi}(\mathbf{x}^*)$ represents an estimation of the density $\bar{\pi}(\mathbf{x})$ at some point $\mathbf{x}^*$. In order to have good performance $\mathbf{x}^*$ should be chosen in high probability region.

• **Computing $\widehat{\pi}(\mathbf{x}^*)$ with C-KDE**. Given a posterior sample, we can substitute $\bar{\pi}(\mathbf{x})$ with its C-KDE of Eq. (6) at a point $\mathbf{x}^*$, so

$$\widehat{Z}_{\text{KDE}} = \frac{\pi(\mathbf{x}^*)}{\widehat{\pi}(\mathbf{x}^*)} = \frac{\pi(\mathbf{x}^*)}{\sum_{i=1}^{C} \bar{w}_i \mathcal{N}(\mathbf{x}^*|\mathbf{m}_i, \mathbf{S}_i)}. \quad (7)$$

Interestingly, if we set $C = 1$ and consider the specific choice $\mathbf{x}^* = \mathbf{m}_1$, then $\widehat{Z}_{\text{KDE}}$ corresponds to the so-called Laplace-Metropolis estimator of Z [19].

• **Computing $\widehat{\pi}(\mathbf{x}^*)$ with Chib's method**. In [20], the authors propose using the MCMC samples of $\bar{\pi}(\mathbf{x})$ to build an estimate of the posterior in one point, $\widehat{\pi}(\mathbf{x}^*) = \widehat{\pi}_{\text{Ch}}(\mathbf{x}^*)$, and plug it in the identity

$$\widehat{Z}_{\text{Ch}} = \frac{\pi(\mathbf{x}^*)}{\widehat{\pi}_{\text{Ch}}(\mathbf{x}^*)}. \quad (8)$$

More specifically, let us consider a MH method with independent proposal density $\varphi(\mathbf{x}) = \varphi(\mathbf{x}|\mathbf{x}_{t-1})$, They suggest estimating $\bar{\pi}(\mathbf{x}^*)$ as

$$\widehat{\pi}_{\text{Ch}}(\mathbf{x}^*) = \frac{\frac{1}{N_1}\sum_{i=i}^{N_1}\alpha(\mathbf{x}_i,\mathbf{x}^*)\varphi(\mathbf{x}^*)}{\frac{1}{N_2}\sum_{i=1}^{N_2}\alpha(\mathbf{x}^*,\mathbf{z}_i)}, \tag{9}$$

where $\alpha(\cdot,\cdot)$ is given in Eq. (4), $\{\mathbf{x}_i\}_{i=1}^{N_1}$ are the MCMC samples, and $\{\mathbf{z}_i\}_{i=1}^{N_2}$ is an i.i.d. sample from $\varphi(\mathbf{x})$.

B. Methods based on Importance sampling (IS)

IS is based on rewriting Eq. (2) as an expected value with respect to (w.r.t.) a normalized importance density $\bar{q}(\mathbf{x})$,

$$Z = \int_{\mathcal{X}} \pi(\mathbf{x})d\mathbf{x} = \int_{\mathcal{X}} \frac{\pi(\mathbf{x})}{\bar{q}(\mathbf{x})}\bar{q}(\mathbf{x})d\mathbf{x} = \mathbb{E}_{\bar{q}}\left[\frac{\pi(\mathbf{x})}{\bar{q}(\mathbf{x})}\right]. \tag{10}$$

Considering we draw M independent samples from proposal $\bar{q}(\mathbf{x})$, $\{\mathbf{z}_i\}_{i=1}^{M}$, the general IS estimator of Z is

$$\begin{aligned}\widehat{Z}_{\text{IS}} &= \frac{1}{M}\sum_{i=1}^{M}\frac{\pi(\mathbf{z}_i)}{\bar{q}(\mathbf{z}_i)},\\ &= \frac{1}{M}\sum_{i=1}^{M}\frac{\ell(\mathbf{y}|\mathbf{z}_i)g(\mathbf{z}_i)}{\bar{q}(\mathbf{z}_i)}, \quad \mathbf{z}_i \sim \bar{q}(\mathbf{x}).\end{aligned} \tag{11}$$

The requirement of $\bar{q}(\mathbf{x})$ is having fatter tails than $\bar{\pi}(\mathbf{x})$. In case we simulate $\{\mathbf{z}_i\}_{i=1}^{M} \sim \bar{q}(\mathbf{x})$ but we are unable to evaluate $\bar{q}(\mathbf{x})$, we need to resort to the self-normalized IS estimator. Consider the unnormalized $q(\mathbf{x}) \propto \bar{q}(\mathbf{x})$, thus

$$\widehat{Z}_{\text{IS}}^{\text{Self}} = \frac{1}{\sum_{i=1}^{M}\frac{f(\mathbf{z}_i)}{q(\mathbf{z}_i)}}\sum_{i=1}^{M}\frac{\pi(\mathbf{z}_i)}{q(\mathbf{z}_i)}, \quad \mathbf{z}_i \sim \bar{q}(\mathbf{x}), \tag{12}$$

where $f(\mathbf{x})$ is also a normalized function, i.e., $\int_{\mathcal{X}} f(\mathbf{x})d\mathbf{x} = 1$. Note that $f(\mathbf{x})$ is never sampled, it plays only the role of an auxiliary density. Different estimators correspond to different choices of $\bar{q}(\mathbf{x})$ or $q(\mathbf{x})$, and $f(\mathbf{x})$.

• **Naive Monte Carlo.** Setting $\bar{q}(\mathbf{x}) = g(\mathbf{x})$, i.e., the prior pdf in Eq. (11), we obtain

$$\widehat{Z}_{\text{naive}} = \frac{1}{M}\sum_{i=1}^{M}\ell(\mathbf{y}|\mathbf{z}_i), \quad \mathbf{z}_i \sim g(\mathbf{x}), \tag{13}$$

namely, an average of the likelihood values of M samples $\mathbf{z}_i$ from the prior density $g(\mathbf{x})$ [21]. This estimator is highly inefficient when the likelihood and the prior functions present probability masses located in different regions of the state space. Note that this method does not require MCMC samples (but only to draw samples from the prior pdf). We discuss here for its simplicity and for the sake of completeness.

• **Harmonic mean estimator.** Setting $\bar{q}(\mathbf{x}) = \bar{\pi}(\mathbf{x})$, so that $q(\mathbf{x}) = \pi(\mathbf{x})$, and $f(\mathbf{x}) = g(\mathbf{x})$ in Eq. (12), we obtain

$$\widehat{Z}_{\text{har}} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)}}, \quad \mathbf{x}_i \sim \bar{\pi}(\mathbf{x}), \tag{14}$$

namely, the harmonic mean of likelihood values of a sample from the posterior $\bar{\pi}(\mathbf{x})$. This estimator may also be derived from the representation $\frac{1}{Z} = \mathbb{E}_{\bar{\pi}}\left[\frac{1}{\ell(\mathbf{y}|\mathbf{x})}\right]$. The harmonic mean estimator has been defined as the "worst" Monte Carlo estimator of Z [22], as it usually exhibits huge variance.

• **Reverse importance sampling (RIS)**. The RIS estimator may be derived in different ways. Formally it is the self-normalized IS estimator in Eq. (12) setting $q(\mathbf{x}) = \pi(\mathbf{x})$ and using a generic auxiliary density $f(\mathbf{x})$, i.e.,

$$\widehat{Z}_{\text{RIS}} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{x}_i)}{\pi(\mathbf{x}_i)}}, \quad \mathbf{x}_i \sim \bar{\pi}(\mathbf{x}). \tag{15}$$

RIS can be also derived from $\frac{1}{Z} = \mathbb{E}_{\bar{\pi}}\left[\frac{f(\mathbf{x})}{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}\right]$, hence is a generalization of the harmonic mean estimator. Unlike in standard IS, $f(\mathbf{x})$ should have lighter tails then $\bar{\pi}(\mathbf{x})$ [3]. In this work, we suggest to use the C-KDE approximation in Eq. (6) as $f(\mathbf{x})$. This choice is advantageous since the C-KDE of $\bar{\pi}(\mathbf{x})$, with $h = 0$, has smaller variance than $\bar{\pi}(\mathbf{x})$ [18].

• **Layered adaptive importance sampling (LAIS)**. The vanilla version of LAIS uses all posterior samples obtained via MCMC to build a standard KDE of $\bar{\pi}(\mathbf{x})$ in Eq. (5), and uses it as proposal pdf $\bar{q}(\mathbf{x})$ for an IS scheme [23]. In this work, we consider the "compressed" version of LAIS (CLAIS), that uses the C-KDE of $\bar{\pi}(\mathbf{x})$ as importance density, i.e., setting $\bar{q}(\mathbf{x})$ to be Eq. (6). Then, drawing $\mathbf{z}_1, ..., \mathbf{z}_M$ from $\bar{q}(\mathbf{x}) = \sum_{i=1}^{C}\bar{w}_i\mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{S}_i)$, the final CLAIS estimator is

$$\widehat{Z}_{\text{CLAIS}} = \frac{1}{M}\sum_{k=1}^{M}\frac{\pi(\mathbf{z}_k)}{\sum_{i=1}^{C}\bar{w}_i\mathcal{N}(\mathbf{z}_k|\mathbf{m}_i,\mathbf{S}_i)}. \tag{16}$$

Recall that all $\mathbf{m}_j$ and $\mathbf{S}_j$ are obtained by a C-KDE using $\mathbf{x}_1, ..., \mathbf{x}_N \sim \bar{\pi}(\mathbf{x})$, generated by an MCMC technique. It is interesting to note that the above estimator can be viewed as an average of Eq. (7), which only uses a single point $\mathbf{x}^*$.

IV. Numerical simulations

A. First numerical experiment

In order to compare the different techniques, we consider the computation of Z of a posterior which is a mixture of two D-dimensional Gaussian densities. We consider a conjugate model with a prior formed by a mixture of two Gaussian and a Gaussian likelihood. Given the observation vector $\mathbf{y}$ (simulated according to the model), we consider a D-dimensional Gaussian likelihood function

$$\ell(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{x}, \mathbf{\Lambda}), \tag{17}$$

with covariance $\mathbf{\Lambda}$, and a D-dimensional Gaussian mixture prior

$$g(\mathbf{x}) = \beta\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\text{pr}}^{(1)}, \mathbf{\Sigma}_{\text{pr}}^{(1)}) + (1-\beta)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\text{pr}}^{(2)}, \mathbf{\Sigma}_{\text{pr}}^{(2)}), \tag{18}$$

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

$$\bar{\pi}(\mathbf{x}) = \alpha\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\text{post}}^{(1)}, \mathbf{\Sigma}_{\text{post}}^{(1)}) + (1-\alpha)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\text{post}}^{(2)}, \mathbf{\Sigma}_{\text{post}}^{(2)}), \tag{19}$$

where the parameters $\alpha \in [0,1]$, $\boldsymbol{\mu}_{\text{post}}^{(i)}$ and $\boldsymbol{\Sigma}_{\text{post}}^{(i)}$ can be obtained in closed-form from β, $\boldsymbol{\mu}_{\text{pr}}^{(i)}$, $\boldsymbol{\Sigma}_{\text{pr}}^{(i)}$ and $\mathbf{y}$. Having the posterior in closed-form allows to compute exactly the marginal likelihood Z. In this case, we can also draw samples directly from the posterior. We can interpret this scenario as the use of an MCMC scheme whose performance is extremely good. We desire to compare the performance of different estimators of Z for different D and different distances between the posterior modes

$$\texttt{dist} = \left\| \boldsymbol{\mu}_{\text{post}}^{(1)} - \boldsymbol{\mu}_{\text{post}}^{(2)} \right\|_2. \tag{20}$$

This distance can be controlled by changing the distance between the prior modes. More specifically, we choose $\boldsymbol{\Sigma} = 50\mathbf{I}_D$, $\boldsymbol{\Sigma}_{\text{pr}}^{(1)} = \boldsymbol{\Sigma}_{\text{pr}}^{(2)} = 30\mathbf{I}_D$, 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; and for the prior means we chose $\boldsymbol{\mu}_{\text{pr}}^{(1)} = -\boldsymbol{\mu}_{\text{pr}}^{(2)} = L\mathbf{1}_D$. We can change the distance between the modes of the prior, $\left\| \boldsymbol{\mu}_{\text{pr}}^{(1)} - \boldsymbol{\mu}_{\text{pr}}^{(2)} \right\|_2 = 2L\sqrt{D}$, and, hence between the modes of the posterior, by varying $L \in \mathbb{R}^+$. Specifically, we select 11 different values $L \in \{1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51\}$ and compare: **(1)** the Naive-MC estimator with 10^4 samples from the prior; **(2)** the HM estimator, **(3)** Laplace estimator and **(4)** RIS with KDE with 10^4 posterior samples; **(5)** a fair application of CLAIS (denoted as CLAIS-f) with only $5 \cdot 10^3$ posterior samples and other $5 \cdot 10^3$ samples from the importance density layer. Recall that RIS (with KDE) and CLAIS-f apply a clustering technique (e.g., k-means algorithm) with $C = 4$ clusters. RIS (with KDE) and CLAIS-f also need to select a bandwidth parameter h. Both methods obtain good results despite of the choice of h. We find that the choices $h = 2$ for RIS and $h = 10$ for CLAIS-f show the average performance of both. We test the techniques in dimension $D = 1$ and $D = 5$. We compute the relative Mean Absolute Error (MAE) in the estimation of Z, averaged over 200 independent simulations.

The results are depicted in Figure 1. They show that RIS and the CLAIS-f achieve the best overall performances. Their relative error remain small and rather constant for all distances, for $D = 1$ and $D = 5$. The RIS estimator performs as well as CLAIS-f in both $D = 1$ and $D = 5$, and even better for small distances in $D = 1$. The lowest relative error corresponds to the Naive MC estimator (for the smallest distance, when prior and posterior are very similar), although it rapidly get outperformed by RIS and CLAIS-f. The Laplace estimator provides poor results since the posterior is bimodal. As one could expect, the estimators that make use of the posterior sample to adapt its importance density, i.e. RIS and CLAIS-f, achieve best performances, being almost independent to increasing the distance between the modes. The HM estimator confirms its reputation of "bad" estimator (see [22], [24])

B. Experiment with real data

Now, we consider the first example in [25], a two-dimensional nonlinear regression problem. The outcome variable Y_i = BOD (mg/L) is modeled in terms of t_i (days; known parameters), as $Y_i = x_1(1 - e^{-x_2 t_i}) + \epsilon_i$, $i = 1, \ldots, 6$, where the ϵ_i's are independent $\mathcal{N}(\epsilon|0, \sigma^2)$ errors (where σ is unknown but we will integrate it out), and the unknown variables of interest are $\mathbf{x} = [x_1, x_2]^\top$. Following [25], we consider the uniform priors $g_1(x_1) = \frac{1}{60}$, $x_1 \in [0, 60]$, and $g_2(x_2) = \frac{1}{6}, x_2 \in [0, 6]$, and an improper prior for σ, $g_3(\sigma) \propto \frac{1}{\sigma}$. The two-dimensional target function $\pi(\mathbf{x}) = \pi(x_1, x_2)$, after integrating out σ, is

$$\pi(\mathbf{x}) = \frac{1}{60}\frac{1}{6}\frac{1}{\pi^3} \frac{8}{\left\{\sum_{i=1}^{6}[y_i - x_1(1 - \exp(-x_2 t_i))]^2\right\}^3},$$

for which we can compute its normalizing constant Z by a costly two dimensional grid. Thus, the ground-truth is $\log Z = -16.208$. We compare the relative MAE of the estimators in (7), (8), (15) and (16). We run a MH algorithm using the prior density as proposal pdf, i.e., $\varphi(\mathbf{x}) = g_1(x_1)g_2(x_2)$ for $N = 10^4$ iterations. We test different numbers of clusters $C \in \{2, 10\}$ and different values of the bandwidth $h = \{0.1, 1, 2, 3, 4, 5\}$. All the results are averaged over 1000 independent runs.

Since CLAIS needs more evaluations of $\pi(\mathbf{x})$, in order to have a fair comparison, we set $N' = N/2$ (where N' is the number of samples in MH) and $M = N/2$. We denote it as fair CLAIS (CLAIS-f) scheme. Moreover, $\widehat{Z}_{\text{KDE}}$ and $\widehat{Z}_{\text{Ch}}$ also

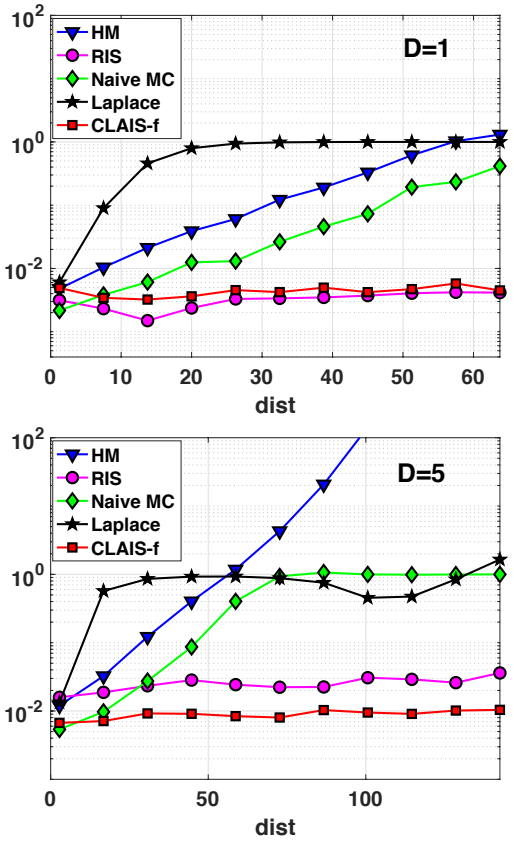

Fig. 1. Relative MAE versus dist; **(Up)** dimension $D = 1$ and **(Down)** dimension $D = 5$.

need specifying the point $\mathbf{x}^*$. We considered two scenarios: (i) using $\mathbf{x}^* = [19, 1]$ that is intentionally located very close to the posterior mode; (ii) using random $\mathbf{x}^*$ drawn from the priors, i.e., from $\varphi(\mathbf{x})$. 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 estimate the relative MAE of $\widehat{Z}_{\text{KDE}}, \widehat{Z}_{\text{KDE-f}}, \widehat{Z}_{\text{CJ}}, \widehat{Z}_{\text{CJ-f}}, \widehat{Z}_{\text{RIS}}$ and $\widehat{Z}_{\text{CLAIS-f}}$, where the "f" stands for "fair". Figure 2 shows the results of the experiment. As expected, the relative error of $\widehat{Z}_{\text{KDE-f}}$ and $\widehat{Z}_{\text{Ch-f}}$ are bigger than those of $\widehat{Z}_{\text{KDE}}$ and $\widehat{Z}_{\text{Ch}}$, respectively. The relative error of $\widehat{Z}_{\text{KDE-f}}$ is huge ($\sim 10^2$) and for that reason, it is impossible to show the curve jointly with the others.

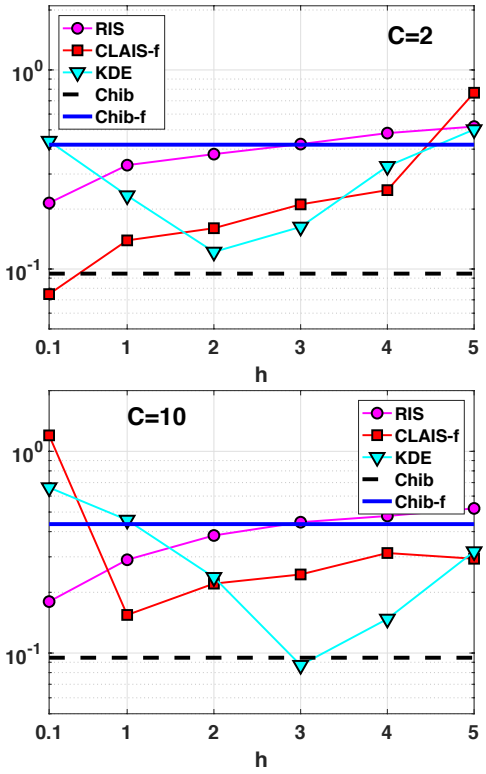

Fig. 2. Relative MAE versus h; **(Up)** $C = 2$ and **(Down)** $C = 10$.

Regarding $\widehat{Z}_{\text{KDE}}$, it shows a great overall performance for all choices of C due to using a $\mathbf{x}^*$ really close to the posterior mode. It also shows that the best choice of h for $\widehat{Z}_{\text{KDE}}$ grows as C increases. The error of $\widehat{Z}_{\text{Ch-f}}$ is around 5 times bigger than its "unfair" counterpart $\widehat{Z}_{\text{Ch}}$, which seems to be the best overall estimator. Nevertheless, the performance of $\widehat{Z}_{\text{Ch-f}}$ provides reasonable results. Regarding $\widehat{Z}_{\text{CLAIS-f}}$ (recall $N' = M = \frac{N}{2}$ for a fair comparison), we observe CLAIS-f needs bigger values of h as C grows. If h is not too small, CLAIS-f provides robust results as function of h. Finally, as we expect, RIS works better with smaller h since the C-KDE pdf, in Eq. (6), will have lighter tails than the posterior. In summary, $\widehat{Z}_{\text{Ch}}$ and $\widehat{Z}_{\text{KDE}}$ can be powerful estimators but the choice of $\mathbf{x}^*$ is critical (especially, for $\widehat{Z}_{\text{KDE}}$ where the results can be catastrophic). $\widehat{Z}_{\text{CLAIS-f}}$ is also a robust alternative that outperforms $\widehat{Z}_{\text{RIS}}$ except that for some values of h.

V. Conclusions

We have described and compared different estimators of the marginal likelihood based on MCMC simulations. We have proposed the use of C-KDE in some of them. Numerical results show that CLAIS and Chib's estimators provide efficient and robust performance.

References

[1] W. J. Fitzgerald, "Markov chain Monte Carlo methods with applications to signal processing," *Signal Processing*, vol. 81, no. 1, pp. 3 – 18, 2001.

[2] F. Llorente, L. Martino, and D. Delgado, "Parallel Metropolis-Hastings coupler," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 953–957, 2019.

[3] C. P. Robert and D. Wraith, "Computational methods for Bayesian model choice," *AIP conference proceedings*, vol. 1193, no. 1, pp. 251–262, 2009.

[4] G.l Stoltz and M. Rousset, "Free energy computations: A mathematical perspective," *World Scientific*, 2010.

[5] M. H. Chen, Q. M. Shao, and J. G. Ibrahim, *Monte Carlo methods in Bayesian computation*, 2012.

[6] N. Friel and J. Wyse, "Estimating the evidence-a review," *Statistica Neerlandica*, vol. 66, no. 3, pp. 288–308, 2012.

[7] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.

[8] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

[9] 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.

[10] J. Skilling, "Nested sampling for general Bayesian computation," *Bayesian analysis*, vol. 1, no. 4, pp. 833–859, 2006.

[11] N. Chopin and C. P. Robert, "Properties of nested sampling," *Biometrika*, vol. 97, no. 3, pp. 741–755, 2010.

[12] N. G. Polson and J. G. Scott, "Vertical-likelihood Monte Carlo," *arXiv preprint arXiv:1409.3601*, 2014.

[13] L. Martino, J. Read, and D. Luengo, "Independent doubly adaptive rejection Metropolis sampling," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7998–8002.

[14] L. Martino, D. Luengo, and J. Míguez, "Independent random sampling methods," *Springer*, 2018.

[15] H. Haario, E. Saksman, and J. Tamminen, "An adaptive Metropolis algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223–242, 2001.

[16] C. P. Robert, V. Elvira, N. Tawn, and C. Wu, "Accelerating MCMC algorithms," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 10, no. 5, pp. e1435, 2018.

[17] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz, "Nonparametric and semiparametric models," 2012.

[18] L. Martino and V. Elvira, "Compressed Monte Carlo for distributed Bayesian inference," *viXra:1811.0505*, 2018.

[19] S. M. Lewis and A. E. Raftery, "Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 648–655, 1997.

[20] 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.

[21] 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.

[22] R. Neal, "The harmonic mean of the likelihood: worst Monte Carlo method ever," *https://radfordneal.wordpress.com/*, 2008.

[23] L. Martino, V. Elvira, D. Luengo, and J. Corander, "Layered adaptive importance sampling," *Statistics and Computing*, vol. 27, no. 3, pp. 599–623, 2017.

[24] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review," *viXra:2001.0052*, 2019.

[25] 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.