Bayesian Updating and Marginal Likelihood Estimation by Cross Entropy based Importance Sampling

Michael Engel*, Oindrila Kanjilal**, Iason Papaioannou, Daniel Straub

Engineering Risk Analysis Group, Technical University of Munich, 80290 Munich

Abstract

We introduce a novel simulation-based method for Bayesian analysis to learn model parameters based on data. The method employs importance sampling (IS) to construct a sample-based approximation of the posterior probability density function (PDF) and estimate the marginal likelihood. We propose to build the IS density through an adaptive sampling approach based on the cross entropy (CE) method. The aim is to identify the parameters of a chosen parametric distribution family that minimize its Kullback-Leibler divergence from the target posterior PDF. An adaptive multi-level approach, based on tempering of the likelihood function, is proposed to efficiently solve this CE optimization problem. We investigate the appropriate choice of the parametric distribution, depending on the number of uncertain model parameters and nature of the posterior density. Numerical studies demonstrate the performance of the proposed method.

Keywords: Bayesian updating, marginal likelihood, importance sampling, cross entropy, tempering, stratified resampling

*Currently, Ph.D. student in the Chair of Remote Sensing Technology, Technical University of Munich, 80290 Munich. Formerly, M.Sc. student in the Engineering Risk Analysis Group, Technical University of Munich, 80290 Munich.

**Corresponding author

Email addresses: `m.engel@tum.de` (Michael Engel), `oindrila.kanjilal@tum.de` (Oindrila Kanjilal), `iason.papaioannou@tum.de` (Iason Papaioannou), `straub@tum.de` (Daniel Straub)

1. Introduction

Learning the parameters of a computational model from measured data is a task arising in many fields of science and engineering. Bayesian analysis provides a systematic framework to address this task by treating the parameters as random variables. The prior knowledge of the random variables, given by the prior joint probability density function (PDF), is updated with data to a posterior PDF through the application of Bayes' rule. The normalizing constant of the posterior PDF is known as marginal likelihood and its evaluation is required in Bayesian model class selection, i.e., to assess the plausibility of each model from a set of available models.

In most practical applications, the posterior PDF does not admit analytical solutions, hence, numerical methods are employed to estimate the posterior distribution, most of which are sampling-based. A popular class of sampling methods is Markov chain Monte Carlo (MCMC) [1]. MCMC samples states of a Markov chain with stationary distribution equal to the target posterior distribution. MCMC approaches have the drawback that it is difficult to assess if the simulated Markov chain has converged and the chain can get stuck in local modes of the posterior. These issues can be partially addressed by sequential Monte Carlo (SMC) samplers [2, 3, 4, 5]. The basic idea of SMC is to sequentially sample a set of intermediate densities that gradually approach the posterior density. In each step, sample generation can be performed with a resample-move scheme [3]. In the resample step, samples drawn from the previous distribution in the sequence are weighted according to the next distribution and then resampled to obtain uniformly weighted samples. This is followed by the move or rejuvenation step, which applies an MCMC algorithm whose target distribution is the current distribution in the sequence. SMC is often combined with MCMC algorithms that are able to efficiently treat high-dimensional priors, including the preconditioned Crank-Nicolson (pCN) sampler and its adaptive variants [6, 7, 8], see, e.g., [9, 10]. In the context of Bayesian analysis, the intermediate densities in SMC are defined by tempering the likelihood function by an exponent [2, 5]. Adaptive SMC approaches determine the tempering parameters and, hence, the intermediate distributions, on the fly during the simulation [11, 12, 13]. An alternative approach, originally proposed in [14], transforms the Bayesian inference problem into an equivalent reliability (rare event) estimation problem, which is then solved using state-of-the-art structural reliability methods. This approach, termed Bayesian Updating with Structural reliability meth-

ods (BUS), has been combined with MCMC-based subset simulation (SuS) in [14, 15, 16]. BUS-SuS is an adaptive simulation method that, similar to SMC, performs a sequence of sampling steps to gradually approach the posterior distribution. An important advantage of SMC and BUS-SuS compared to MCMC sampling is that both methods return an estimate of the marginal likelihood as a by-product of the simulation.

Both SMC and BUS-SuS employ MCMC methods to generate samples from the intermediate distributions, which inevitably leads to dependent posterior samples. Alternatively, one can employ importance sampling (IS) to obtain a weighted sample approximation of the posterior distribution through sampling from a suitable IS density [17]. The same samples can be used to obtain an estimate of the marginal likelihood. The effectiveness of IS strongly depends on the choice of the IS density. Iterative IS methods adapt the IS density through a sequence of sampling steps. Each batch of intermediate samples is used to identify the next IS density through fitting mixture distribution models or through kernel density estimation [18, 19]. Proper sample initialization is a key aspect of these approaches. Often the first set of samples is generated with MCMC methods or with the help of Laplace approximations [20, 21].

The cross entropy (CE) method is an adaptive IS method that constructs the IS density through fitting a parametric distribution model [22]. The parameters are determined through minimizing the Kullback-Leibler (KL) divergence between a target distribution and the parametric model. The CE method was originally proposed for rare event estimation [23]. In this context, the target density is the optimal IS density for estimating the probability of the rare event. The CE optimization problem is solved through a sequential sampling approach designed such that the samples gradually approach regions of high probability mass of the optimal IS density. Such multi-level CE-based IS methods have been extensively studied in the field of rare event estimation and reliability analysis [23, 24, 25, 26, 27, 28, 29, 30, 31], while the rationale of the multi-level CE approach has also been applied to the solution of optimization problems [32, 25]. In the context of Bayesian analysis, the CE approach has been applied for marginal likelihood estimation in [33]. This method employs samples from the posterior obtained, e.g., with MCMC to solve the CE optimization problem in a manner similar to the method of [34].

This paper proposes a new CE-based IS method for Bayesian Updating (CEBU). The method applies the multi-level CE approach to fit a parametric

3

distribution family such that it best approximates the posterior distribution. Similar to the approach of [28], the adaptive sampling process is performed through defining a sequence of intermediate densities that starting from the prior gradually approach the posterior density. This density sequence is defined by a tempering of the likelihood, as in classical SMC approaches [2, 5]. The tempering parameters are determined in an adaptive manner, following the adaptive SMC paradigm [11, 12]. Hence, the proposed approach is completely adaptive and does not require an initial choice of the distribution parameters, e.g., based on an initial MCMC sampling step, as is the case in iterative IS methods. We investigate two distribution families as parametric densities, the Gaussian mixture (GM) [26, 27] and the von-Mises-Fisher-Nakagami mixture (vMFNM) [28]. Using mixture models enables one to efficiently describe multimodal posterior distributions.

The outline of the paper is as follows. We start by giving the fundamental definitions of Bayesian inverse problems and the IS method in Sec. 2. In Sec. 3, we derive the proposed method by applying the CE method to Bayesian updating using tempering. The choice of the parametric IS density is described in Sec. 4. In Sec. 5 we demonstrate the performance of the method by means of three numerical examples and compare it to that of the adaptive SMC and BUS-SuS methods.

2. Background

This section provides a brief background to Bayesian inverse problems and importance sampling.

2.1. Bayesian inverse problem

We consider a model of an underlying process represented by a mapping $G : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the space of parameters of the model and $\mathcal{Y}$ denotes the space of model outcomes. The model G may come, e.g., from a system of ordinary differential equations, partial differential equations or expert knowledge. We assume that both the parameter space and the outcome space are finite-dimensional, precisely $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^d$. Consider now that a set of measurements $\boldsymbol{y} \in \mathcal{Y}$ of the underlying process is available. In inverse problems, the objective is to identify the system parameter $\boldsymbol{\theta}$ given the measurements $\boldsymbol{y}$.

In Bayesian inverse problems, the system parameters are assumed to be uncertain, and one describes the prior belief on the parameter vector $\boldsymbol{\theta} \in \mathcal{X}$

by PDF $p_0 : \mathcal{X} \to \mathbb{R}_{\geq 0}$. Given system measurements $\boldsymbol{y}$, the formal objective is to estimate the posterior PDF $p_{\boldsymbol{y}} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ for the parameters. We note that we use a notation similar to [35]. The posterior PDF is obtained by application of Bayes' rule as

$$p_{\boldsymbol{y}}(\boldsymbol{\theta}) := \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta}) \cdot p_0(\boldsymbol{\theta})}{Z_{\boldsymbol{y}}}, \tag{1}$$

where the likelihood function $L_{\boldsymbol{y}} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ describes how likely the data (collection of measurements) is for a particular instance $\boldsymbol{\theta}$ of the system parameters. The normalizing constant $Z_{\boldsymbol{y}}$ is given by

$$Z_{\boldsymbol{y}} := \int_{\mathcal{X}} L_{\boldsymbol{y}}(\boldsymbol{\theta}) p_0(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

and is referred to as the marginal likelihood or model evidence. It is a measure for the plausibility of the model G, and is required for model class selection and model averaging [36].

Evaluation of the posterior PDF analytically is seldom possible, and one has to resort to numerical approaches. In Sec. 3 we introduce an importance sampling method based on the principle of CE minimization to approximate the posterior PDF.

2.2. Importance sampling

Before presenting the proposed CE-based IS method, we provide a brief background on IS. Let $\psi(\boldsymbol{\theta}) = L_{\boldsymbol{y}}(\boldsymbol{\theta}) \cdot p_0(\boldsymbol{\theta})$ be the unnormalized posterior PDF. Consider the generic problem of evaluating the expected value of a function $H : \mathcal{X} \to \mathbb{R}$ with respect to $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. One can express this expectation as

$$\begin{aligned} I = \mathrm{E}_{p_{\boldsymbol{y}}}[H(\boldsymbol{\theta})] &= \frac{1}{Z_{\boldsymbol{y}}} \int_{\mathcal{X}} H(\boldsymbol{\theta}) \psi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{Z_{\boldsymbol{y}}} \int_{\mathcal{X}} H(\boldsymbol{\theta}) W(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{3}$$

where $\mathrm{E}_{p_{\boldsymbol{y}}}[\cdot]$ denotes expectation with respect to $p_{\boldsymbol{y}}(\boldsymbol{\theta})$, $h : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is the IS density, satisfying $\mathrm{supp}(h) \supseteq \mathrm{supp}(\psi)$, and $W(\boldsymbol{\theta}) = \frac{\psi(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}$ is the importance

weight function. In a similar way, the marginal likelihood $Z_{\boldsymbol{y}}$ can be expressed as

$$Z_{\boldsymbol{y}} = \int_{\mathcal{X}} W(\boldsymbol{\theta})h(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}. \tag{4}$$

Eqs. 3 and 4 give rise to the following self-normalized IS estimator for I:

$$\hat{I} = \frac{1}{N}\frac{1}{\hat{Z}_{\boldsymbol{y}}}\sum_{i=1}^{N} W(\boldsymbol{\theta}_i)H(\boldsymbol{\theta}_i), \tag{5}$$

where $\{\boldsymbol{\theta}_i, i = 1, \ldots, N\}$ are independently drawn samples from $h(\boldsymbol{\theta})$ and $\hat{Z}_{\boldsymbol{y}}$ is the estimate of the marginal likelihood, given by

$$\hat{Z}_{\boldsymbol{y}} = \frac{1}{N}\sum_{i=1}^{N} W(\boldsymbol{\theta}_i). \tag{6}$$

The estimator of Eq. 5 corresponds to the expectation with respect to a Monte Carlo approximation of the PDF $p_{\boldsymbol{y}}(\boldsymbol{\theta})$, given by

$$p_{\boldsymbol{y}}(\boldsymbol{\theta}) \approx \frac{1}{N\hat{Z}_{\boldsymbol{y}}}\sum_{i=1}^{N} W(\boldsymbol{\theta}_i)\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_i), \tag{7}$$

where $\delta(\cdot)$ is the Dirac function. Often the interest is in estimating Eq. 3 for a wide range of functions H. In such case, it is a good idea to choose an IS density h that is similar to $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. In fact, $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ would be the optimal choice of the IS density for estimating the marginal likelihood $Z_{\boldsymbol{y}}$, since it leads to a zero-variance estimator $\hat{Z}_{\boldsymbol{y}}$.

3. Cross entropy method for Bayesian updating

In this section, we present a novel method to determine the posterior PDF of the uncertain system parameters. The proposed method employs IS and is based on the principle of CE minimization. The CE method is an adaptive IS method that determines a near-optimal approximation of a target PDF

6

by minimizing the KL divergence between the target density and a chosen family of parametric distributions. It was originally proposed for estimating rare event probabilities in [23] and was later extended for solving optimization problems in [32]. In the context of Bayesian inverse problems, we implement the CE method to approximate the posterior PDF $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. Hence, our target density of interest is the PDF $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. The fitted parametric PDF can then be used to estimate the marginal likelihood through IS according to Eq. 6. Furthermore, following Eq. 7, a sample-based approximation of the posterior PDF can be obtained.

Let $h(\boldsymbol{\theta}, \boldsymbol{\nu})$ be a family of parametric densities, where $\boldsymbol{\nu} \in \mathcal{V}$ is the parameter vector, which contains the prior PDF $p_0(\boldsymbol{\theta})$. The KL divergence between $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta}, \boldsymbol{\nu})$ is a measure of difference between the two PDFs and is defined as [22]

$$
\begin{aligned}
D_{KL}\left(p_{\boldsymbol{y}}(\boldsymbol{\theta}) \| h(\boldsymbol{\theta}, \boldsymbol{\nu})\right) &= \mathrm{E}_{p_{\boldsymbol{y}}}\left[\log\left(\frac{p_{\boldsymbol{y}}(\boldsymbol{\theta})}{h(\boldsymbol{\theta}, \boldsymbol{\nu})}\right)\right] \\
&= \int_{\mathcal{X}} \log\left(p_{\boldsymbol{y}}(\boldsymbol{\theta})\right) p_{\boldsymbol{y}}(\boldsymbol{\theta}) d\boldsymbol{\theta} - \int_{\mathcal{X}} \log\left(h(\boldsymbol{\theta}, \boldsymbol{\nu})\right) p_{\boldsymbol{y}}(\boldsymbol{\theta}) d\boldsymbol{\theta}.
\end{aligned}
\tag{8}
$$

The KL divergence can be interpreted as the information lost when approximating $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ with $h(\boldsymbol{\theta}, \boldsymbol{\nu})$ [37]. It holds $D_{KL}\left(p_{\boldsymbol{y}}(\boldsymbol{\theta}) \| h(\boldsymbol{\theta}, \boldsymbol{\nu})\right) \geq 0$, with $D_{KL}\left(p_{\boldsymbol{y}}(\boldsymbol{\theta}) \| h(\boldsymbol{\theta}, \boldsymbol{\nu})\right) = 0$ if and only if $h(\boldsymbol{\theta}, \boldsymbol{\nu}) = p_{\boldsymbol{y}}(\boldsymbol{\theta})$.

The basic idea of the CE method is to determine the parameter vector $\boldsymbol{\nu}^*$ that minimizes the KL divergence of Eq. 8. The PDF $h(\boldsymbol{\theta}, \boldsymbol{\nu}^*)$ is then a close approximation of the posterior PDF $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. As the parametric density appears only in the second term in Eq. 8, a minimum of the KL divergence can be found by maximizing only this second term. Additionally, substituting Eq. 1 for $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ into Eq. 8 results in the following optimization problem:

$$
\begin{aligned}
\boldsymbol{\nu}^* &= \arg\min_{\boldsymbol{\nu} \in \mathcal{V}} D_{KL}\left(p_{\boldsymbol{y}}(\boldsymbol{\theta}) \| h(\boldsymbol{\theta}, \boldsymbol{\nu})\right) \\
&= \arg\max_{\boldsymbol{\nu} \in \mathcal{V}} \int_{\mathcal{X}} L_{\boldsymbol{y}}(\boldsymbol{\theta}) p_0(\boldsymbol{\theta}) \log(h(\boldsymbol{\theta}, \boldsymbol{\nu})) d\boldsymbol{\theta} \\
&= \arg\max_{\boldsymbol{\nu} \in \mathcal{V}} \mathrm{E}_{p_0}[L_{\boldsymbol{y}}(\boldsymbol{\theta}) \log(h(\boldsymbol{\theta}, \boldsymbol{\nu}))].
\end{aligned}
\tag{9}
$$

The expectation in Eq. 9 can be approximated using a set of samples $\{\boldsymbol{\theta}_k; k = 1, \ldots, N_S\}$ from $p_0(\boldsymbol{\theta})$, which gives the sample-counterpart of the CE optimization problem:

$$\boldsymbol{\nu}^* = \arg\max_{\nu \in \mathcal{V}} \left\{ \frac{1}{N_S} \sum_{i=1}^{N_S} L_{\boldsymbol{y}}(\boldsymbol{\theta}_i) \log\left(h(\boldsymbol{\theta}_i, \boldsymbol{\nu})\right) \right\}. \tag{10}$$

Often, the objective functions in Eqs. 9 and 10 are convex and differentiable with respect to $\boldsymbol{\nu}$, and therefore the problem is solved by setting the gradient of the objective function equal to zero. In order to obtain a good sample approximation of Eq. 9 with Eq. 10, a considerable number of the samples drawn from $p_0(\boldsymbol{\theta})$ should lie in the high probability region of $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. Ensuring this is computationally challenging when the data is highly informative, i.e., when $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ differs significantly from $p_0(\boldsymbol{\theta})$. To overcome this difficulty, we develop a multi-level approach to solve the CE optimization problem.

3.1. Multi-level cross entropy method based on tempering

Consider a sequence of intermediate densities $\{f_t(\boldsymbol{\theta}); t = 0, \ldots, T\}$, which start from the prior PDF and gradually approach the posterior PDF, i.e., $f_0(\boldsymbol{\theta}) := p_0(\boldsymbol{\theta})$ and $f_T(\boldsymbol{\theta}) := p_{\boldsymbol{y}}(\boldsymbol{\theta})$. Such a distribution sequence can be constructed by tempering the likelihood function:

$$f_t(\boldsymbol{\theta}) := \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta})^{\beta_t} p_0(\boldsymbol{\theta})}{Z_t}, \tag{11}$$

where $\{\beta_t; t = 0, \ldots, T\}$ is the sequence of tempering parameters, which satisfy $0 = \beta_0 < \beta_1 < \ldots < \beta_t < \ldots < \beta_{T-1} < \beta_T = 1$ and Z_t is the normalizing constant of $f_t(\boldsymbol{\theta})$. The tempering parameters bridge the gap between $p_0(\boldsymbol{\theta})$ and $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ by allowing a smooth transition between the two PDFs. The change between two consecutive densities in the sequence can be made small through an appropriate selection of $\{\beta_t; t = 0, \ldots, T\}$. Consequently, the region over which $f_t(\boldsymbol{\theta})$ is significant can be adequately represented by a small number of samples drawn from $f_{t-1}(\boldsymbol{\theta})$, or a close approximation of $f_{t-1}(\boldsymbol{\theta})$. An illustration of the sequence of Eq. 11 is shown in Figure 1. This distribution sequence has been previously used to develop MCMC-based SMC samplers for Bayesian analysis [2, 5, 11].

In the multi-level CE method, we solve the CE optimization problem sequentially for the intermediate densities defined in Eq. 11. This leads to a sequence of parameter vectors $\{\boldsymbol{\nu}_t; t = 1, \ldots, T\}$. The goal is to find a final parameter vector $\boldsymbol{\nu}_T$ close to the optimal parameter $\boldsymbol{\nu}^*$. The parameter

vector $\boldsymbol{\nu}_t$ is determined by minimizing the KL divergence between $f_t(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta}, \boldsymbol{\nu})$, which results in the following optimization problem:

$$\boldsymbol{\nu}_t = \arg\max_{\boldsymbol{\nu} \in \mathcal{V}} \mathrm{E}_{f_t}[\log(h(\boldsymbol{\theta}, \boldsymbol{\nu}))]. \tag{12}$$

We estimate the expectation in Eq. 12 by IS using a set of samples from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$, where $\hat{\boldsymbol{\nu}}_{t-1}$ is the solution of the optimization problem in the previous step. After substituting the expression of $f_t(\boldsymbol{\theta})$ into Eq. 12, the sample counter-part of the above CE optimization problem is obtained as

$$\hat{\boldsymbol{\nu}}_t = \arg\max_{\boldsymbol{\nu} \in \mathcal{V}} \left\{ \frac{1}{N_S} \sum_{i=1}^{N_S} W_t(\boldsymbol{\theta}_i, \hat{\boldsymbol{\nu}}_{t-1}) \log\left(h(\boldsymbol{\theta}_i, \boldsymbol{\nu})\right) \right\}, \tag{13}$$

where $W_t(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1}) = \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta})^{\beta_t} p_0(\boldsymbol{\theta})}{h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})}$ and $\{\boldsymbol{\theta}_i; i = 1, \ldots, N_S\}$ are independent samples drawn from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$. A default choice of $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_0)$ is the prior PDF $p_0(\boldsymbol{\theta})$.

3.2. Adaptive selection of tempering parameter and convergence

The accuracy and computational efficiency of the multi-level CE method depends crucially on the choice of the tempering parameters. The difference between the values of $\{\beta_t; t = 0, \ldots, T\}$ directly influences the change between the respective intermediate densities. In order to get a good sample approximation in Eq. 13 with a limited number of samples, the intermediate PDF $f_t(\boldsymbol{\theta})$ should not differ substantially from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$, which is a close approximation of the PDF $f_{t-1}(\boldsymbol{\theta})$. We ensure this by selecting the tempering parameters adaptively using a criterion based on the effective sample size (ESS). For N_S samples drawn from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$, the ESS gives the equivalent number of proposed samples from $f_t(\boldsymbol{\theta})$ available for fitting the parametric density. The ESS is therefore a measure of difference between the target density $f_t(\boldsymbol{\theta})$ and the sampling density $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$, and is defined as [17, 38, 39]

$$\mathrm{ESS}_t(\beta_t) = \frac{N_S}{1 + \hat{\delta}_{W_t}^2(\beta_t)}, \tag{14}$$

where $\hat{\delta}_{W_t}(\beta_t)$ denotes the sample coefficient of variation of the weights $\{W_t(\boldsymbol{\theta}_i, \hat{\boldsymbol{\nu}}_{t-1}); i = 1, \ldots, N_S\}$. We choose β_t such that the ESS in Eq. 14 is equal to some predefined target value $\mathrm{ESS}_{target} > 0$. Thus, we define

9
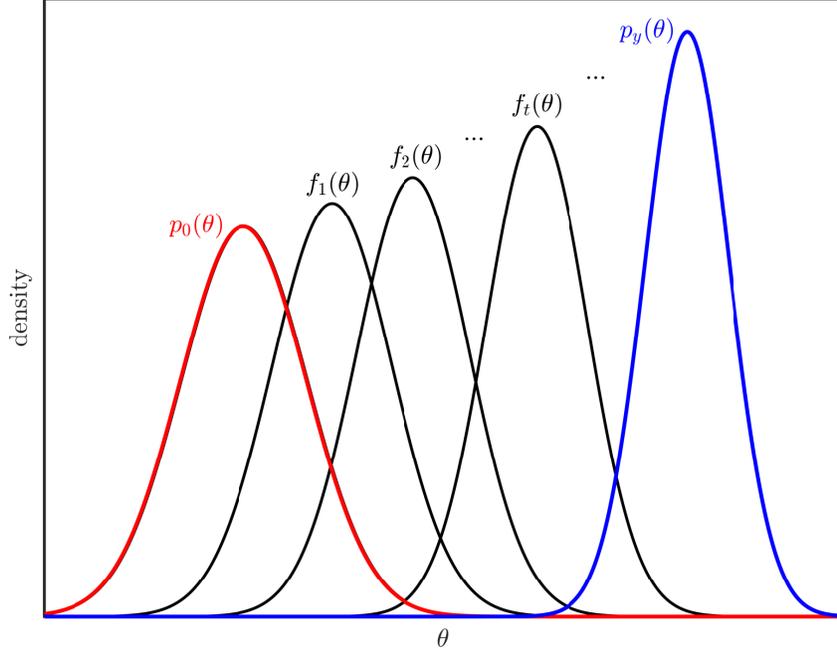
Figure 1: Sequence of intermediate target densities $f_t(\theta)$ from prior $p_0(\theta)$ to posterior $p_{\boldsymbol{y}}(\theta)$

$$\beta_t = \underset{\beta \in (\beta_{t-1}, 1]}{\arg\min} \left\{ (\text{ESS}_{target} - \text{ESS}_t(\beta))^2 \right\}. \tag{15}$$

This implies that the tempering parameter is increased such that the number of proposed samples is equivalent to a chosen ESS in each level. Instead of directly prescribing ESS_{target}, the user can specify the number of samples per step N_S and the desired coefficient of variation of the importance weights δ_{target}, and determine the corresponding ESS_{target} by analogy with Eq. 14.

We now make the assumption that the parametric density $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$ is equal to $f_{t-1}(\boldsymbol{\theta})$, in which case the importance weight $W_t(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$ in Eq. 13 is proportional to the lumped importance weight

$$\widetilde{W}_t(\boldsymbol{\theta}) = L(\boldsymbol{\theta})^{\beta_t - \beta_{t-1}}. \tag{16}$$

After substituting $\widetilde{W}_t(\boldsymbol{\theta})$ for $W_t(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$ in Eq. 14, and using the fundamental relation $\mathrm{Var}[\bullet] = \mathrm{E}[\bullet^2] - \mathrm{E}[\bullet]^2$, the tempering parameter in Eq. 15 is given by

$$\beta_t = \underset{\beta \in (\beta_{t-1}, 1]}{\arg\min} \left\{ \left(\frac{N_S}{1 + \delta_{target}^2} - \frac{\left(\sum_{i=1}^{N_S} L_{\boldsymbol{y}}(\boldsymbol{\theta}_i)^{(\beta - \beta_{t-1})} \right)^2}{\sum_{i=1}^{N_S} L_{\boldsymbol{y}}(\boldsymbol{\theta}_i)^{(2\beta - 2\beta_{t-1})}} \right)^2 \right\}, \qquad (17)$$

where $\{\boldsymbol{\theta}_i; i = 1, \ldots, N_S\}$ are independent samples generated from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_{t-1})$. The modified selection criteria in Eq. 17 does not depend on the parameters $\{\boldsymbol{\nu}_t; t = 0, \ldots, T-1\}$. This provides improved robustness in the convergence of the algorithm in problems where the parametric density is not able to perfectly describe the target density. The adaptive procedure is stopped when the value of the tempering parameter determined based on Eq. 17 is equal to 1. After convergence, at the T-th step, the final parameter vector $\hat{\boldsymbol{\nu}}_T$ is determined by solving the optimization problem in Eq. 13 with $\beta_T = 1$.

3.3. Marginal likelihood

The density $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$ determined by the CE method can be used to estimate the marginal likelihood $Z_{\boldsymbol{y}}$ by IS. As mentioned in Sec. 2.2, the posterior PDF is the optimal IS density for estimating $Z_{\boldsymbol{y}}$. Hence, $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$, being a close approximation of the posterior, is expected to be a good IS density for estimating $Z_{\boldsymbol{y}}$, provided that the chosen parametric model is sufficiently flexible. Following Eq. 4, the marginal likelihood can be expressed as

$$\begin{aligned} Z_{\boldsymbol{y}} &= \int_{\mathcal{X}} L_{\boldsymbol{y}}(\boldsymbol{\theta}) p_0(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\mathcal{X}} W_T(\boldsymbol{\theta}) h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T) d\boldsymbol{\theta}. \end{aligned} \qquad (18)$$

Eq. 18 leads to the following IS estimator

$$\hat{Z}_{\boldsymbol{y}} = \frac{1}{N} \sum_{i=1}^{N} W_{Ti}, \qquad (19)$$

where $W_{Ti} = \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta}_i) p_0(\boldsymbol{\theta}_i)}{h(\boldsymbol{\theta}_i, \hat{\boldsymbol{\nu}}_T)}$ and $\{\boldsymbol{\theta}_i, i = 1, \ldots, N\}$ are independent samples generated from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$. The number of samples N in Eq. 19 can be chosen

11

freely and can be different from the number of samples per level N_S used to solve the CE optimization problem.

The variance of the estimator $\hat{Z}_{\boldsymbol{y}}$ can be estimated using the samples $\{\boldsymbol{\theta}_i, i = 1, \ldots, N\}$, as follows [17]

$$\hat{\sigma}^2_{\hat{Z}_{\boldsymbol{y}}} = \frac{1}{N} \widehat{\mathrm{Var}}(W_{Ti}), \tag{20}$$

where $\widehat{\mathrm{Var}}(W_{Ti})$ is the sample variance of the weights $\{W_{Ti}, i = 1, \ldots, N\}$. The samples $\{\boldsymbol{\theta}_i, i = 1, \ldots, N\}$ can be used to estimate expectations of arbitrary function with respect to the posterior, using the self-normalizing IS estimator of Eq. 5. Estimation of the variance of such estimates is discussed, e.g., in [17].

3.4. Generation of posterior samples through resampling

The density $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$ determined by the CE method is an approximation of the posterior PDF $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ for the chosen parametric family. Using a set of independent samples $\{\boldsymbol{\theta}_i, i = 1, \ldots, N\}$ drawn from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$, we obtain the following IS approximation of the posterior PDF

$$p_{\boldsymbol{y}}(\boldsymbol{\theta}) \approx \frac{1}{N\hat{Z}_{\boldsymbol{y}}} \sum_{i=1}^{N} W_{Ti} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_i), \tag{21}$$

where W_{Ti} are defined as in Eq. 19. To obtain independent and identically distributed (i.i.d.) samples from the density in Eq. 21, one can apply a resampling algorithm, such as stratified resampling, multinomial resampling or residual resampling [40, 41]. In the present study, we employ a global stratified resampling scheme adapted from [42]. This approach is summarized in Algorithm 1. We note that an alternative approach for generating samples from the posterior PDF is to employ the density $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$ as a proposal density in the independent Metropolis-Hastings algorithm [43]. The full CE-based IS algorithm for Bayesian updating (CEBU) is summarized in Algorithm 2.

4. Choice of parametric density in the cross entropy method

In the CE method, the family of parametric distributions $h(\boldsymbol{\theta}, \boldsymbol{\nu})$ is typically chosen such that it contains the nominal density of the uncertain model

---
Algorithm 1: Stratified Resampling

Result: N samples following the posterior PDF $p_{\boldsymbol{y}}(\boldsymbol{\theta})$ given N samples from the final importance sampling (IS) density $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$ and their corresponding importance weights.

Input: Independent samples $\{\boldsymbol{\theta}_i; i = 1, \ldots, N\}$ from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$
Importance weights $\{W_{Ti}, i = 1, \ldots, N\}$

Main Algorithm:
Compute the cumulative sum of the normalized importance weights
$\overline{W}_{Ti} = \dfrac{\sum_{k=1}^{i} W_{Tk}}{\sum_{j=1}^{N} W_{Tj}}$. Draw uniformly distributed samples
$w_i \sim U\left(\dfrac{i-1}{N}, \dfrac{i}{N}\right)$, for $i = 1, \ldots, N$.

Set $i = 1$ and $j = 1$
repeat
    if $w_i < \overline{W}_{Tj}$ **then**
        Set $\widetilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_j$.
        Increase $i = i + 1$.
    else
        Increase $j = j + 1$.
until $i = N$

Output: Samples $\{\widetilde{\boldsymbol{\theta}}_i; i = 1, \ldots, N\}$ distributed according to $p_{\boldsymbol{y}}(\boldsymbol{\theta})$.
---

parameters. In the context of Bayesian inference, the nominal density corresponds to the prior PDF $p_0(\boldsymbol{\theta})$. Without loss of generality, we represent $p_0(\boldsymbol{\theta})$ in terms of an underlying outcome space $\boldsymbol{u} \in \mathbb{R}^n$ of independent standard normal random variables. To this end, we employ a mapping $\boldsymbol{u} = T_{ISO}(\boldsymbol{\theta})$ from the outcome space of the prior PDF $p_0(\boldsymbol{\theta})$ to the outcome space of the standard normal PDF, where T_{ISO} is a one-to-one iso-probabilistic transformation [44, 45]. For the specific case of independent priors, i.e., if $p_0(\boldsymbol{\theta})$ can be written as $p_0(\boldsymbol{\theta}) = \prod_{k=1}^{n} p_0^k\left(\theta^{(k)}\right)$, T_{ISO} is comprised of a set of marginal transformations given by $u^{(k)} = \Phi^{-1}\left(P_0^k\left(\theta^{(k)}\right)\right); k = 1, \ldots, n$, where $u^{(k)}$ and $\theta^{(k)}$, respectively, denote the k-th components of the vectors $\boldsymbol{u}$ and $\boldsymbol{\theta}$,

Algorithm 2: Cross entropy based importance sampling for Bayesian updating (**CEBU**)

Result: Estimate of the marginal likelihood, $Z_{\boldsymbol{y}}$, and generate a set of samples from a posterior distribution $p_{\boldsymbol{y}}(\boldsymbol{\theta})$

Input: Likelihood function, $L_{\boldsymbol{y}}(\boldsymbol{\theta})$
Prior distribution of the model parameters, $p_0(\boldsymbol{\theta})$
Choice of the parametric density $h(\boldsymbol{\theta}, \boldsymbol{\nu})$
Sample size for CE optimization, N_S
Target CoV of the weights in each level, δ_{target}
Desired number of samples, N, from the posterior distribution

Algorithm:
Set $t = 0$ and $\beta_0 = 0$.
Select $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_0)$ as the prior PDF $p_0(\boldsymbol{\theta})$.
repeat

1. Draw N_S samples $\{\boldsymbol{\theta}_i; i = 1, \ldots, N_S\}$ from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_t)$.

2. Evaluate the likelihood function values $\{L_{\boldsymbol{y}}(\boldsymbol{\theta}_i); i = 1; \ldots, N_S\}$.

3. Determine the tempering parameter β_{t+1} by solving the optimization problem in Eq. 17.

4. Compute the weights $\{W_t(\boldsymbol{\theta}_i, \hat{\boldsymbol{\nu}}_t); i = 1, \ldots, N_S\}$.

5. Determine $\hat{\boldsymbol{\nu}}_{t+1}$ by solving the CE optimization problem in Eq. 13 using the weights evaluated in the previous step.

6. Set $t = t + 1$.

until $\beta_t = 1$

7. Draw N samples $\{\boldsymbol{\theta}_i; i = 1, \ldots, N\}$ from $h(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}_T)$ and compute $\{W_{Ti} = \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta}_i)p_0(\boldsymbol{\theta}_i)}{h(\boldsymbol{\theta}_i, \hat{\boldsymbol{\nu}}_T)}; i = 1, \ldots, N\}$.

8. Compute an estimate of the marginal likelihood by taking the sample mean of $\{W_{Ti}; i = 1, \ldots, N\}$: $\hat{Z}_{\boldsymbol{y}} = \frac{1}{N} \sum_{i=1}^{N} W_{Ti}$.

9. Resample $\{\boldsymbol{\theta}_i; i = 1, \ldots, N\}$ according to Algorithm 1 to obtain samples from the posterior PDF.

14

$\Phi(\cdot)$ is the standard Gaussian cumulative distribution function (CDF) and $P_0^k(\cdot)$ is the marginal prior CDF of the k-th component. For the case of dependent priors, T_{ISO} is given by the Nataf transformation [44], which considers a Gaussian copula for the dependence structure, or, more generally, the Rosenblatt transformation [45]. The posterior PDF in the transformed space is given by:

$$\phi_{\boldsymbol{y}}(\boldsymbol{u}) = \frac{L_{\boldsymbol{y}}(T_{ISO}^{-1}(\boldsymbol{u}))\phi_0(\boldsymbol{u})}{Z_{\boldsymbol{y}}}, \tag{22}$$

where $\phi_0(\boldsymbol{u})$ is the n-dimensional standard normal joint PDF. We note that the normalizing constant in Eq. 22 is the same as the one in Eq. 1 because of the iso-probabilistic nature of the transformation, as shown in Appendix A.

The updating problem is solved in the transformed space. This approach, which was also employed in [13], has certain advantages. In particular, the uncertainty in the prior distribution is normalized and its support is unbounded. In the context of the CE method, the transformation T_{ISO} facilitates the choice of the parametric distribution family for a wide range of Bayesian updating problems.

To implement the CE method, we consider two different parametric families $h(\boldsymbol{u}, \boldsymbol{\nu})$, presented in Secs. 4.1 and 4.2, which contain $\phi_0(\boldsymbol{u})$. We determine the parameter vectors $\{\hat{\boldsymbol{\nu}}_t, t = 1, \ldots, T\}$ by sequentially minimizing the KL divergence between $h(\boldsymbol{u}, \boldsymbol{\nu})$ and the intermediate target densities in the standard normal space according to the procedure in Algorithm 2. The marginal likelihood is directly estimated using Eq. 19 with the weights evaluated at the final step. Samples $\{\boldsymbol{u}_i, i = 1, \ldots, N\}$ drawn from the final density $h(\boldsymbol{u}, \hat{\boldsymbol{\nu}}_T)$ are transformed to equivalent samples $\{\boldsymbol{\theta}_i, i = 1, \ldots, N\}$ in the original parameter space by applying the inverse iso-probabilistic transformation $\boldsymbol{\theta}_i = T_{ISO}^{-1}(\boldsymbol{u}_i)$. These samples are then resampled according to Algorithm 1 to obtain samples distributed according to $p_{\boldsymbol{y}}(\boldsymbol{\theta})$. We note that the transformation to the original space can alternatively be applied following resampling in the transformed space.

We consider two different distribution families for application of the CE method in standard normal space. The first is the GM distribution, which allows efficient representation of multi-modal posterior distributions. The second is a distribution model, which enables efficient application of the method to certain high-dimensional problems.

15

4.1. Gaussian mixture

The GM density is defined as the weighted sum of K multivariate normal PDFs:

$$f_{GM}(\boldsymbol{u}, [\underline{\mu}, \underline{\Sigma}, \underline{\alpha}]) = \sum_{k=1}^{K} \alpha_k f_{\mathcal{N}}(\boldsymbol{u}, [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)), \tag{23}$$

where $\underline{\mu}$ is the set of K mean vectors, $\underline{\Sigma}$ is the set of K covariance matrices, $\underline{\alpha}$ is the set of K weights for the mixture terms summing up to one and

$$f_{\mathcal{N}}(\boldsymbol{u}, [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]) = \frac{1}{\sqrt{2^n \pi^n \det \boldsymbol{\Sigma}_k}} \exp\left(-\frac{1}{2}\left((\boldsymbol{u} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{u} - \boldsymbol{\mu}_k)\right)\right). \tag{24}$$

The total number of parameters to be fitted in the GM model is $Kn(n+3)/2 + (K-1)$. For the uni-modal case, i.e., $K = 1$, closed-form expressions for the parameter update in Eq. 13 are available, e.g., in [22]. For the multi-modal case, we apply the Expectation-Maximization algorithm suggested in [27], with initialization of the sample allocation to the different modes through application of k-means clustering, to update the distribution parameters.

4.2. von-Mises-Fisher-Nakagami mixture

It is well-known that the probability mass of a high-dimensional PDF with i.i.d. marginals concentrates on an important ring of radius proportional to $\sqrt{n}$ [46, 47]. This is commonly known as the concentration of norm phenomenon. In the case of mixture distribution with i.i.d. density components, we expect that in finite dimensions the probability mass concentrates around multiple important rings. This is demonstrated in Figure 2 for a GM model.

As a consequence, if the posterior PDF is a mixture of high-dimensional i.i.d. densities, then it can be efficiently characterized by a distribution in polar coordinates. The polar coordinate representation of a sample $\boldsymbol{u} \in \mathbb{R}^n$ is $\boldsymbol{u} = r \cdot \boldsymbol{a}$, where $\boldsymbol{a} = \boldsymbol{u}/\|\boldsymbol{u}\|$ is the direction and $r = \|\boldsymbol{u}\|$ is the radius. In [28], a mixture distribution model in polar coordinates was introduced in the context of the CE method for rare event simulation, which is shown to perform well in high dimensional problems. The model consists of mixture

components that combine the von-Mises-Fisher (vMF) distribution for the direction $\boldsymbol{a}$ with the Nakagami distribution for the radius r and is, hence, termed the von-Mises-Fisher-Nakagami (vMFN) mixture.
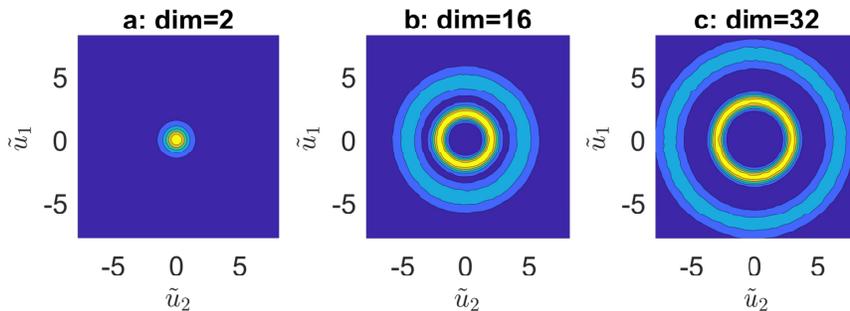
Figure 2: Illustration of the concentration of norm phenomenon for a GM model with two modes both mean zero, a variance of 0.3 and 1.5 and a weighting of 0.35 and 0.65, respectively. The figure shows this GM density in an underlying 2-dimensional space $[\tilde{u}_1, \tilde{u}_2]$. a: Contour plot of 2-dimensional GM density- b: Contour plot of 16-dimensional GM density - c: Contour plot of 32-dimensional GM density

The vMFN PDF is defined in polar coordinates as:

$$
\begin{aligned}
f_{vMFN}&(r, \boldsymbol{a}, [\boldsymbol{\mu}, \kappa, \Omega, m]) \\
&= f_{vMF}(\boldsymbol{a}, [\boldsymbol{\mu}, \kappa]) \cdot f_N(r, [\Omega, m]) \\
&= \frac{\kappa^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)} e^{\kappa \boldsymbol{\mu}^T \boldsymbol{a}} \cdot \frac{2m^m}{\Gamma(m)\Omega^m} r^{2m-1} e^{-\frac{m}{\Omega} r^2}.
\end{aligned}
\tag{25}
$$

where f_{vMF} is the PDF of the vMF distribution and f_N is the PDF of the Nakagami distribution. The concentration parameter κ describes how concentrated the samples are around the mean direction $\boldsymbol{\mu}$. This means that for increasing κ the samples will be closer to the mean direction whereas for $\kappa = 0$ they are uniformly distributed over the unit hypersphere. Regarding the radius, $m \geq 0.5$ is the shape parameter and $\Omega > 0$ the spread parameter. In Eq. 25, I_k denotes the modified Bessel function of first kind of order k. It is noted that the vMFN distribution reduces to the standard normal distribution for $\kappa = 0$, $\Omega = n$, $m = n/2$ and an arbitrary mean direction $\boldsymbol{\mu}$. The von-Mises-Fisher-Nakagami mixture (vMFNM) density is defined as the weighted sum of K vMFN PDFs:

$$f_{vMFNM}(r, a, [\underline{\mu}, \underline{\kappa}, \underline{\Omega}, \underline{m}, \underline{\alpha}]) = \sum_{k=1}^{K} \alpha_k f_{vMFN}(r, a, [\boldsymbol{\mu}_k, \kappa_k, \Omega_k, m_k]), \quad (26)$$

where $\underline{\mu}$ is the set of K mean directions, $\underline{\kappa}$ the set of K concentration parameters, $\underline{\Omega}$ the set of K radial spread parameters and $\underline{m}$ the set of K radial shape parameters. The total number of parameters in the vMFNM model is $K(n+3)+(K-1)$. This number scales linearly with the dimension n, unlike the number of parameters in the GM model, which scales quadratically with n. Hence, the vMFNM model is easier to fit in high dimensions than the GM model. The parameter update in Eq. 13 is performed by an Expectation Maximization algorithm, as suggested in [28].

5. Numerical investigation

To illustrate the proposed method, three numerical examples are studied: one with a bimodal posterior, and two inverse analyses involving a structural dynamics model and an epidemic model. The CE method is implemented using both GM and vMFNM models, described in Sec. 4. The two resulting approaches are abbreviated as CEBU-GM and CEBU-vMFNM. The number of samples drawn from the posterior is always set equal to the number of samples per level during CE optimization, i.e., $N = N_S$. In each example, the results from the proposed method are compared to those from the adaptive variant of the SuS-based BUS approach (aBUS) [16] and the adaptive SMC method [12]. For aBUS, the intermediate conditional probability is set to 0.1. The adaptive SMC approach is implemented using the approach discussed in Sec. 3.2 to evaluate the tempering parameters. Both aBUS and SMC are implemented in the standard normal space following the approach discussed in Sec. 4. The MCMC algorithm for performing the move step in aBUS and SMC is an adaptive pCN sampler, also known as adaptive conditional sampling algorithm [6, 7].

The algorithms are run 100 times for all examples to obtain statistics of the estimates. For the case where a reference solution cannot be evaluated analytically, the reference is obtained through a modified rejection sampling algorithm [48, 49] that allows generating approximately i.i.d. posterior samples without prior knowledge of the maximum of the likelihood function.

To assess the quality of the CE-based posterior approximation of Eq. 21, we employ the normalized effective sample size (nESS) [18]

$$\text{nESS} = \frac{\text{ESS}}{N} = \frac{1}{1 + \hat{\delta}^2_{W_T}}, \tag{27}$$

where ESS is the effective sample size introduced in Eq. 14, N the number of proposed samples and $\hat{\delta}_{W_T}$ the coefficient of variation of the importance weights of the proposed samples. The nESS is a measure of the goodness of fit of the parametric IS density and takes values in $(0, 1]$. The larger the nESS, the smaller the variance of the resulting sample-based estimators.

5.1. Bimodal posterior

The first example involves the approximation of a bimodal posterior density, previously studied in [11, 13, 50]. The aim of this example is to investigate whether CEBU recognizes the bimodal structure of the posterior and how its performance scales with the dimension of the sample space. Multimodal posteriors are challenging for standard random walk MCMC-based approaches as the generated chains can get stuck in one of the modes. The likelihood function is modelled by the mixture of two Gaussian densities:

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = \alpha \cdot \mathcal{N}\left(\boldsymbol{\theta}, [\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1]\right) + (1 - \alpha) \cdot \mathcal{N}\left(\boldsymbol{\theta}, [\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2]\right). \tag{28}$$

The fraction α denotes the portion of the probability mass assigned to the first mode. The prior is uniformly distributed on the hypercube $[-2, 2]^n$:

$$p_0(\boldsymbol{\theta}) = \prod_{i=1}^{n} U(\theta_i, [-2, 2]), \tag{29}$$

where $U(\bullet, [a, b])$ denotes the uniform distribution in $[a, b]$.

The mean vectors of the two Gaussians in Eq. 28 are $\boldsymbol{\mu}_1 = 0.5\ \mathbf{1}_n$ and $\boldsymbol{\mu}_2 = -0.5\ \mathbf{1}_n$ with $\mathbf{1}_n$ denoting the n-dimensional vector of ones. We consider two different scenarios for the covariance matrices. In the first, it is $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2 \boldsymbol{I}_n$, with $\sigma = 0.1$ and $\boldsymbol{I}_n$ denoting the identity matrix. This choice of σ leads to a separation of the two posterior modes [11]. In the second scenario, the two covariance matrices are chosen as $\boldsymbol{\Sigma}_1 = \sigma^2 \boldsymbol{R}_1$ and $\boldsymbol{\Sigma}_2 = \sigma^2 \boldsymbol{R}_2$, with

$$R_1 = \begin{bmatrix} 1 & 0.8 & 0 & \cdots \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & & & \ddots \end{bmatrix}$$

and

$$R_2 = \begin{bmatrix} 1 & -0.8 & 0 & \cdots \\ -0.8 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & & & \ddots \end{bmatrix}$$

The latter scenario is of particular interest for assessing the performance of the vMFNM model, which has limited capability of representing dependence.

The dimension n is varied between 2 and 32 in steps of 3, i.e., $n = 2, 5, 8, ..., 32$. The fraction is chosen as $\alpha = 0.9$. The number of samples per level is set to $N_S = 3000$ and the target coefficient of variation of the weights is $\delta_{target} = 1$, corresponding to $\text{ESS}_{target} = 1500$. Since the posterior is bimodal, two-component GM and vMFNM distributions are chosen.

In order to assess the ability of the method to correctly identify the two distinct modes of the posterior, we compute the estimate $\hat{\alpha}$ of the portion of probability mass of the first posterior mode. $\hat{\alpha}$ is estimated by a hard clustering approach using the two Gaussian mixture modes, under the assumption that the uniform prior has no influence on the posterior.

Fig. 3 shows the contours of the estimated PDFs of the fraction $\hat{\alpha}$ for the first and second scenarios in function of the dimension n. The PDFs are estimated with 100 repeated simulation runs. The reference solution is denoted by red color.

CEBU-GM

CEBU-vMFNM

aBUS

SMC
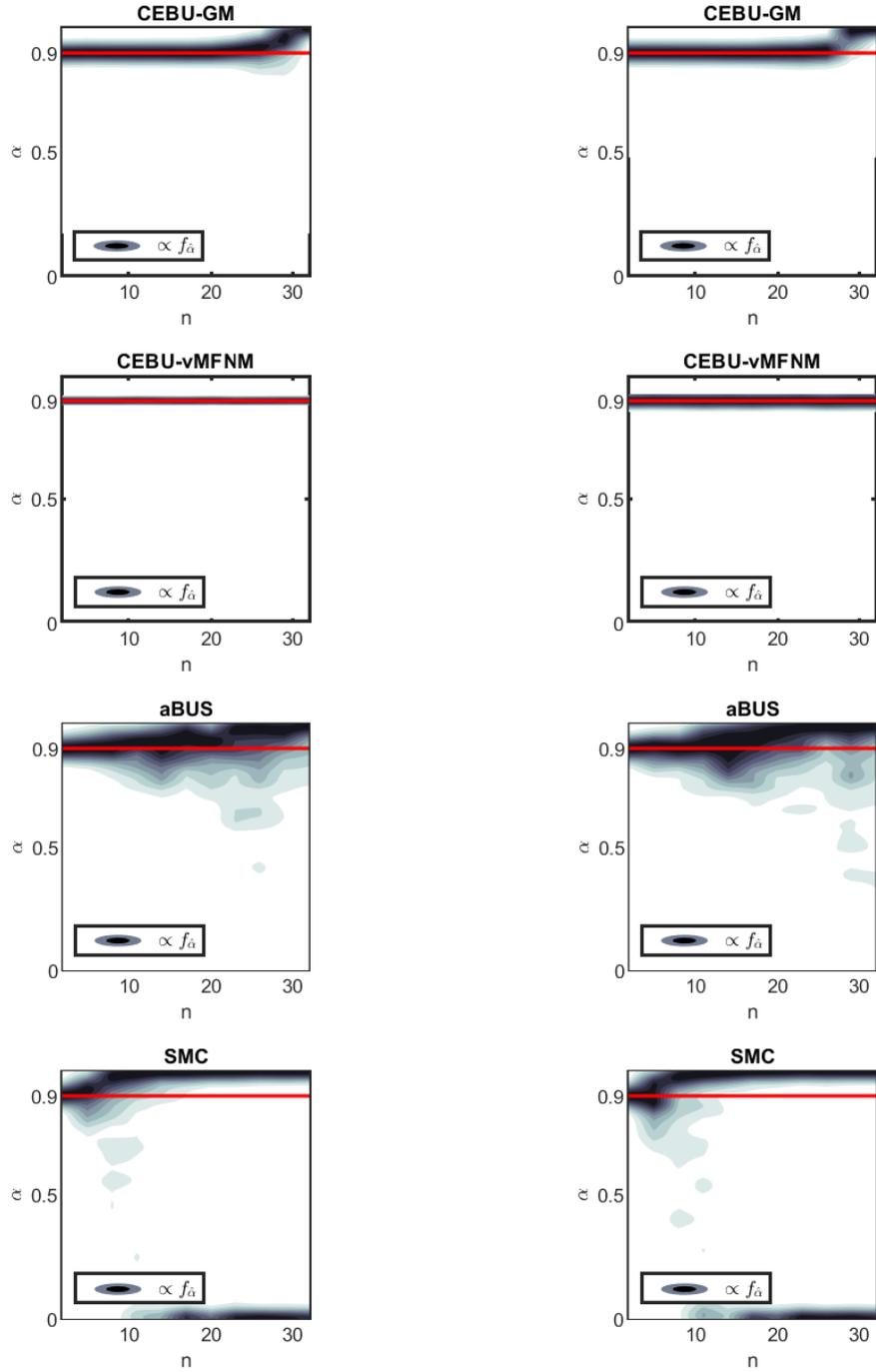
(a)   (b)

Figure 3: a: PDF of the fraction estimate of the bimodal posterior, without dependence, in example 5.1. - b: PDF of the fraction estimate of the bimodal posterior, with dependence, in example 5.1.

The probability mass of the fraction estimates PDFs, obtained by CEBU-vMFNM is concentrated around the true value for all tested dimensions and scenarios. The variability of the estimate is higher for the second scenario, owing to the limited ability of the vMFNM model to capture the parameter dependence. The PDF obtained by CEBU-GM shows higher variability compared to that with the vMFNM model. This is related to the higher dimensionality of the parameter space of the GM model, which requires a larger sample size to obtain confident estimates. The parameter dimension scales quadratically with the dimension of the sample space n for the GM model, as opposed to that of the vMFNM model, which scales linearly with n. This results in a strongly skewed PDF of the fraction estimates for the GM model as the parameter space dimension n becomes larger than 30. For $n = 32$, the number of parameters of the GM model is 1121, which is close to the number of effective samples per level used to fit the parameters (for $N_S = 3000$ and $\delta_{target} = 1$, the target ESS is 1500). On the other hand, the dependence of the mixture components is not an issue for the GM model, as the model is fully capable of representing this dependence. Both the GM- and vMFNM-based CEBU perform significantly better than the aBUS and SMC approaches for this example for $n < 30$. aBUS shows significantly larger variability in the fraction estimates compared to CEBU, although the influence of the sample space dimension is minor due to the properties of implemented pCN sampler. The SMC estimates show a bimodal behavior for $n > 8$, as the algorithm tends to recognize either the first or the second mode of the bimodal posterior. This is attributed to the random walk behavior of the employed MCMC algorithm.

The behavior of the CEBU method with the two considered distribution models is further studied in Fig. 4, which plots the obtained nESS as a function of the sample space dimension n for the two considered scenarios. The vMFNM model shows a remarkable behavior for the first scenario with i.i.d. mixture components as nESS remains close to 1 independent of the dimension. In the second scenario, the obtained nESS is significantly lower, which reflects the low flexibility of the model and indicates that the vMFNM model is not suitable for problems with high correlation coefficients. The behavior of the GM model is similar for both cases; nESS is close to 1 for $n < 10$ and decreases with increase of n due to the quadratic dependence of the model parameters with n.

Fig. 5 shows the total model evaluations (TME) for the different considered approaches and the two different scenarios. Fig. 5a shows that the
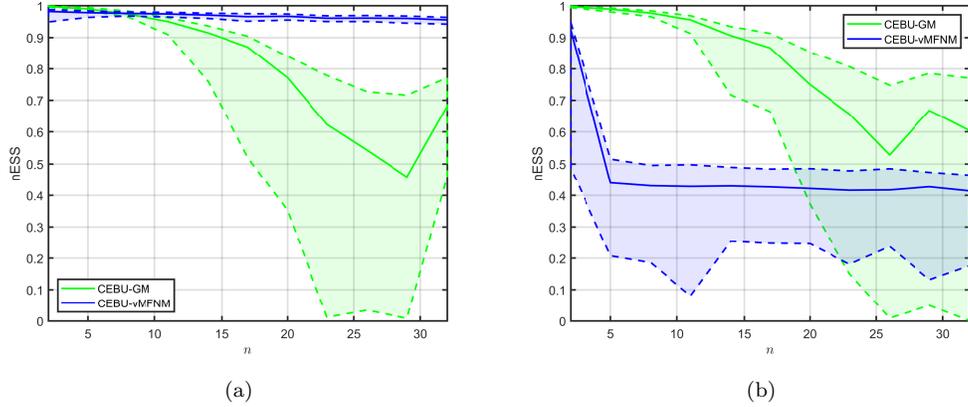
Figure 4: Normalized effective sample size obtained with the vMFNM and GM models for the bimodal posterior in example 5.1, with (a) and without dependence (b). The continuous lines show the medians and the shaded areas show the 95% credible intervals obtained from 100 independent simulation runs.
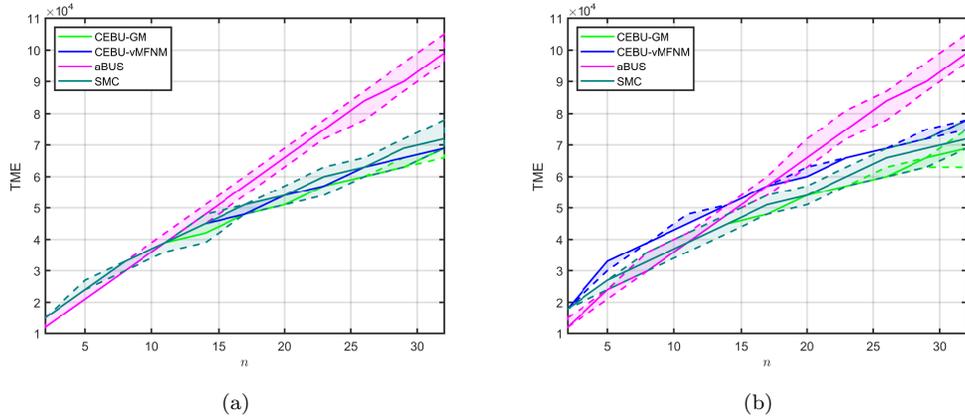
Figure 5: Total model evaluations for the bimodal posterior in example 5.1, with (a) and without dependence (b). The continuous lines show the medians and the shaded areas show the 95% credible intervals obtained from 100 independent simulation runs.

CEBU and SMC approaches lead to similar TME. This is to be expected as both approaches are based on the same definition of intermediate target distributions; CEBU samples from a parametric IS density, while SMC generates samples from the target distributions with MCMC. Comparing Fig. 5a with Fig. 5b, the vMFNM model leads to higher TME in the second scenario, as it requires a larger number of steps to converge to the target ESS. The

TME of aBUS are above all others for $n > 17$ in both scenarios.

5.2. Two degree-of-freedom structure

The second example was originally investigated in [11] and consists of a two-storey building shown in Fig. 6a. The purpose of this example is to investigate the ability of the CEBU method to estimate posterior distributions with strong nonlinear dependence. The structure is modeled as a 2 degree of freedom (DOF) mass-spring-dashpot system. The system is subjected to a narrow-banded ground acceleration such that only the first eigenmode is excited. The ground acceleration $\boldsymbol{a}_{m,g}$ and the roof acceleration $\boldsymbol{a}_{m,r}$ are measured with a frequency of 50Hz for a period of 1s, which results in 50 pairs of measurements. The measurements $\boldsymbol{a}_{m,g}$ are assumed to be perfect, whereas $\boldsymbol{a}_{m,r}$ are assumed to be contaminated with Gaussian white noise and unknown noise variance σ_n^2. The masses of the 2-DOF system are assumed to be known ($m_1 = m_2 = 1$). The uncertain parameters of the model that are inferred from the measurements are the two stiffness parameters k_1 and k_2, the damping ratio ξ, which is assumed to be the same for both DOFs, and the variance of the noise σ_n^2. The likelihood is defined as

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = \exp\left(-\frac{d}{2}\ln(\sigma_n^2) - \frac{1}{2}\frac{\|\boldsymbol{a}_r(k_1, k_2, \xi, \boldsymbol{a}_{m,g}) - \boldsymbol{a}_{m,r}\|_2^2}{\sigma_n^2}\right), \quad (30)$$

where $\boldsymbol{a}_r(k_1, k_2, \xi, \boldsymbol{a}_{m,g})$ is the simulated acceleration at the roof, given the excitation $\boldsymbol{a}_{m,g}$ at the ground, and $d = 50$ denotes the number of measurements. This is solved using the Newmark method (e.g., [51]) with the same time discretization interval as the frequency of available measurements. Uncorrelated uniform priors are used for all random variables:

$$p_{0,k_1}(k) = p_{0,k_2}(k) = U(k, [0, 3000]), \quad (31)$$
$$p_{0,\xi}(\xi) = U(\xi, [0.01, 0.05]), \quad (32)$$
$$p_{0,\sigma_n^2}(\sigma_n^2) = U(\sigma_n^2, [0, 1]). \quad (33)$$

The measurements of the roof acceleration are generated using the model with the true parameter values: $k_1 = k_2 = 1000$, $\xi = 0.03$ and $\sigma_n^2 = 0.2$. The reference posterior is obtained using rejection sampling similar to [49] with 10^5 samples. The resulting joint posterior of the stiffness parameters k_1 and k_2 has a strongly nonlinear dependence, as depicted in Fig. 6b.

24
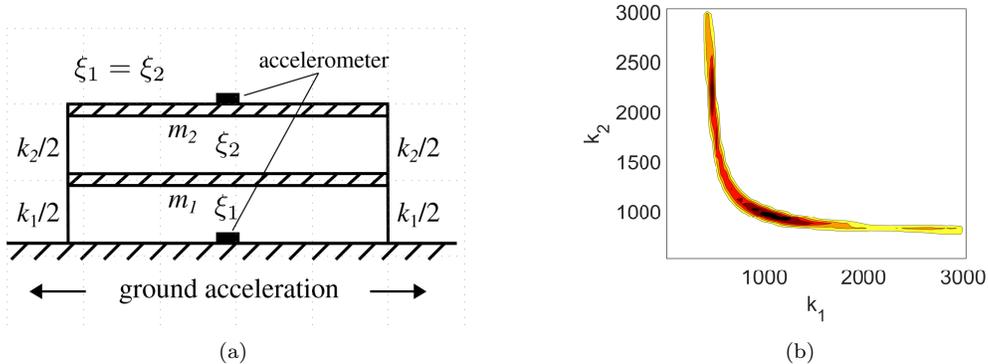
Figure 6: a: Two story building subjected to ground acceleration [11]. - b: Sample based contour plot of the joint PDF of the stiffness parameters k_1 and k_2.

The CEBU method is run with $N_S = 3000$ samples per level and target coefficient of variation of the weights set to $\delta_{target} = 1$. Figs. 7 and 8 show the fitted mixture IS densities for the two stiffness parameters, k_1 and k_2, obtained with different number of mixture terms K. Fig. 7 shows that the GM model leads to a visually adequate representation of the posterior distribution already with $K = 5$ mixture terms, whereas $K = 10$ results in an almost perfect match. In contrast, the vMFNM model is not able to accurately represent the posterior, even with $K = 10$ mixture terms, as shown in Fig. 8. Although the performance of the model appears to increase with increase of the number of mixture terms, its limited flexibility does not allow to capture the full characteristics of the posterior density. This is related to the fact that the individual mixture components of the vMFNM cannot capture a strong dependence between the parameters.

The performance of the two models is further investigated by considering the nESS as a function of the number of mixture terms, which is plotted in Fig. 9. The plot shows that the nESS of the GM model increases as K increases and appears to attain a constant value for $K > 8$. The obtained nESS is rather high, reflecting the high flexibility of the GM model. The situation is different for the vMFNM model. Thereby, the nESS shows a slow increase with K and takes a small value for $K = 10$, which further indicates the limitation of the vMFNM model to capture complex dependence.

Fig. 10 shows the behavior of the relative error of the log-marginal likelihood estimated with the two distribution models. The GM model results in accurate estimates even for $K = 3$, with 95% credible bounds of the er-
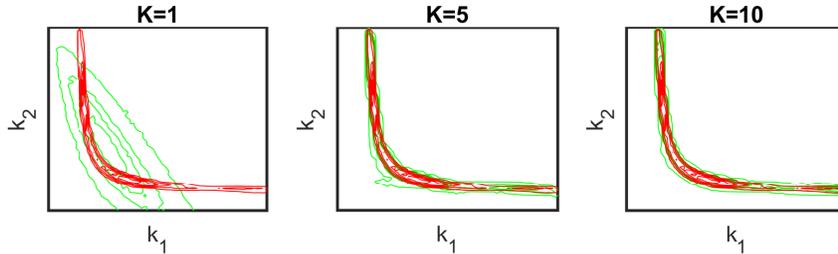
Figure 7: Sample based contour plot of the importance sampling density of the stiffness parameters k_1 and k_2, obtained with the GM distribution with different number of mixture terms, K. Contours of the true posterior distribution are plotted in red.
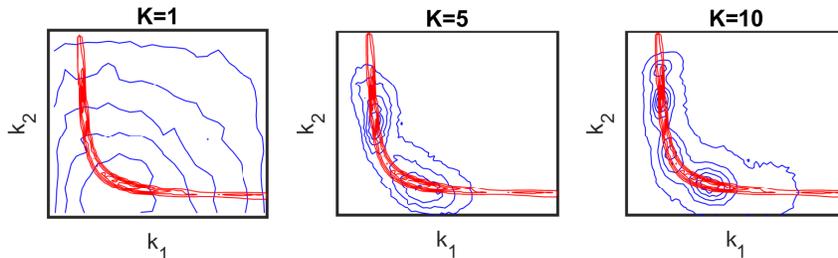
Figure 8: Sample based contour plot of the importance sampling density of the stiffness parameters k_1 and k_2, obtained with the vMFNM distribution with different number of mixture terms, K. Contours of the true posterior distribution are plotted in red.

ror smaller than 0.05. These bounds become smaller with increase of K. The vMFNM model results in skewed estimates of the log-marginal likelihood with the median being about 5% off, and wide credible intervals of the relative error that appear to remain constant for $K > 5$.

In Tab. 1, we compare the estimated statistics (mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$) of the posterior distribution of the model parameters obtained by the CEBU method with $K = 10$ with the ones obtained with aBUS and SMC methods. It can be seen that CEBU-GM, aBUS and SMC result in accurate estimates of the mean of the parameters. A small underestimation is observed in the estimates from CEBU-vMFNM. The associated uncertainty (coefficient of variation, c.o.v.) obtained by CEBU-GM is significantly lower
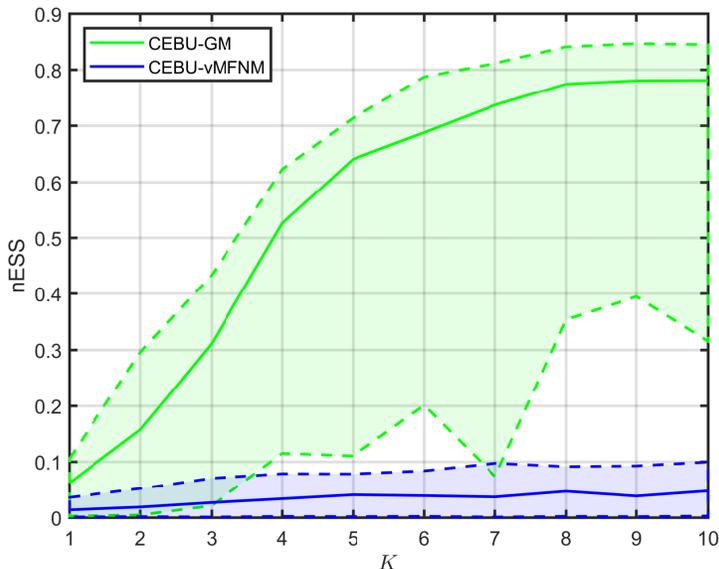
Figure 9: Normalized effective sample size obtained with the GM and vMFNM distributions with different number of mixture terms, K, in example 5.2. The continuous lines show the medians and the shaded areas show the 95% credible intervals obtained from 100 independent simulation runs.

(close to one order of magnitude) compared to both aBUS and SMC. The c.o.v.'s of the mean estimates obtained by CEBU-vMFNM are higher than those of aBUS/SMC. Similar observations can be made for the estimates of the standard deviation of the parameters. The table also compares the mean and standard deviation (std) of the estimated log-marginal likelihood obtained by the different approaches. Again, CEBU-GM results in significantly lower standard deviation than both aBUS and SMC. The high standard deviation obtained by the CEBU-vMFNM method is attributed to one outlier that is observed during the 100 independent runs. As with example 5.1, CEBU-GM and SMC lead to similar TME. The aBUS method requires about 2500 less TME, indicating that the method converges on average in approximately one simulation level faster than CEBU-GM/SMC. The TME of CEBU-vMFNM is significantly higher than the other methods. This is due to the strict requirement of the target c.o.v. of the weights (or, equivalently, the target ESS), which results in about double as many levels for convergence

compared to CEBU-GM/SMC.
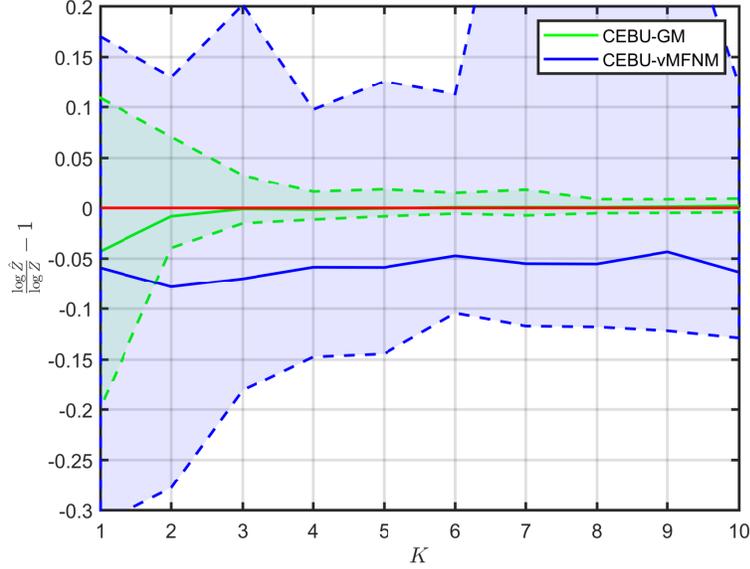
Figure 10: Log-marginal likelihood error obtained with the GM and vMFNM models with different number of mixture terms, K, in example 5.2. The continuous lines show the medians and the shaded areas show the 95% credible intervals obtained from 100 independent simulation runs.

Table 1: Estimates of the posterior statistics of the model parameters, the log-marginal likelihood and the total model evaluations for all investigated methods in example 5.2 . The c.o.v./std of the estimates is obtained from 100 independent simulation runs.

	Reference	CEBU-GM (K=10)	CEBU-vMFNM (K=10)	aBUS	SMC
$\hat{\mu}_{k_1}$ (c.o.v.)	9.48e+02	9.45e+02 (1.44e-02)	8.91e+02 (1.98e-01)	9.40e+02 (9.32e-02)	9.53e+02 (9.50e-02)
$\hat{\mu}_{k_2}$ (c.o.v.)	1.45e+03	1.44e+03 (1.06e-02)	1.40e+03 (1.30e-01)	1.46e+03 (8.29e-02)	1.43e+03 (8.16e-02)
$\hat{\mu}_{\xi}$ (c.o.v.)	3.05e-02	3.05e-02 (7.85e-03)	3.10e-02 (8.51e-02)	3.03e-02 (4.27e-02)	3.06e-02 (5.28e-02)
$\hat{\mu}_{\sigma_n^2}$ (c.o.v.)	2.83e-01	2.82e-01 (6.14e-03)	2.79e-01 (3.88e-02)	2.83e-01 (2.13e-02)	2.84e-01 (3.06e-02)
$\hat{\sigma}_{k_1}$ (c.o.v.)	5.86e+02	5.82e+02 (2.53e-02)	4.16e+02 (2.46e-01)	5.67e+02 (1.40e-01)	5.68e+02 (1.52e-01)
$\hat{\sigma}_{k_2}$ (c.o.v.)	6.31e+02	6.26e+02 (1.81e-02)	5.29e+02 (1.66e-01)	6.24e+02 (8.42e-02)	6.12e+02 (1.13e-01)
$\hat{\sigma}_{\xi}$ (c.o.v.)	9.50e-03	9.51e-03 (1.33e-02)	8.13e-03 (1.64e-01)	9.43e-03 (7.89e-02)	9.37e-03 (8.99e-02)
$\hat{\sigma}_{\sigma_n^2}$ (c.o.v.)	6.08e-02	6.02e-02 (3.25e-02)	5.72e-02 (1.65e-01)	6.07e-02 (6.81e-02)	6.19e-02 (6.69e-02)
$\log \hat{Z}$ (std.)	3.16e+00	3.17e+00 (1.13e-02)	2.76e+00 (2.96e+00)	3.16e+00 (1.57e-01)	3.11e+00 (2.20e-01)
TME (c.o.v.)	17049961	18210 (4.22e-02)	47430 (8.27e-02)	15750 (8.29e-02)	18030 (1.66e-02)

5.3. Epidemiologic model

We apply CEBU to a real-life problem concerning the evolution of the coronavirus pandemic (COVID-19) in Germany. The objective of this example is to test the method on a real data-set. An extended SIR model with a single federal intervention [52, 53] is considered, which describes the evolution of the susceptible (S), infected (I), and removed (R) population [53]:

$$\frac{dS(t)}{dt} = -\beta(t)\frac{I(t)}{N_p}S(t) \tag{34}$$

$$\frac{dI(t)}{dt} = \beta(t)\frac{I(t)}{N_p}S(t) - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma I(t)$$

$$S(0) = N_p - I_0, \ I(0) = I_0, \ R(0) = 0$$

In Eq. 34, β is the unknown infection rate that quantifies the progress of the disease, γ is the removal rate (including recovery and mortality), I_0 is the initial number of infected individuals and N_p is the size of the population. To model an intervention or a series of measures taken by the government, β is considered to be a piece-wise linear function of time [52]:

$$\beta(t) = \begin{cases} \beta_0 & t \leq t_{\text{int}} - \frac{\tau_{\text{int}}}{2} \\ \text{linear} & t_{\text{int}} - \frac{\tau_{\text{int}}}{2} < t < t_{\text{int}} + \frac{\tau_{\text{int}}}{2} \\ k_{\text{int}}\beta_0 & t_{\text{int}} + \frac{\tau_{\text{int}}}{2} \geq t \end{cases}, \tag{35}$$

where $\beta_0 \geq 0$ denotes the basic infection rate, $k_{\text{int}} \geq 0$ is the reduction factor, $t_{\text{int}} \geq 0$ is the intervention time and $\tau_{\text{int}} \geq 0$ is the duration of transition. During $t \in \left(t_{\text{int}} - \frac{\tau_{\text{int}}}{2}, t_{\text{int}} + \frac{\tau_{\text{int}}}{2}\right)$ a linear transition from β_0 to $k_{\text{int}}\beta_0$ is assumed.

The SIR model is calibrated using data $\hat{I}$ of daily confirmed cases reported by the Johns Hopkins University [54]. The analysis in this paper considers data from 1 March to 5 June 2020. The population size for Germany is taken as $N_p = 83.2 \times 10^6$. The likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N_p} \mathrm{NB} \left(\hat{I}_i, \left[\mu_{NB,i}, \sigma_{NB,i}^2 \right] \right), \tag{36}$$

where $\hat{I}_i$ is the measured number of infections at day t_i, following a Negative Binomial (NB) distribution with mean $\mu_{NB,i} = S(t_{i-1}) - S(t_i)$ and variance $\sigma_{NB,i}^2 = \mu_{NB,i} + \frac{\mu_{NB,i}^2}{r}$. Here r denotes the dispersion coefficient. We assume uninformative priors for the model parameters $\boldsymbol{\theta} = (\beta_0, \gamma, t_{\mathrm{int}}, \tau_{\mathrm{int}}, k_{\mathrm{int}}, r)$, i.e., the parameters are taken to be independent with lower and upper bounds given by

$$p_{0,\beta_0}(\beta_0) = U(\beta_0, [0.1, 0.6]), \tag{37}$$
$$p_{0,\gamma}(\gamma) = U(\gamma, [0.07, 0.5]), \tag{38}$$
$$p_{t_{\mathrm{int}}}(t_{\mathrm{int}}) = U(t_{\mathrm{int}}, [1, 97]), \tag{39}$$
$$p_{0,\tau_{\mathrm{int}}}(\tau_{\mathrm{int}}) = U(\tau_{\mathrm{int}}, [1, 14]), \tag{40}$$
$$p_{0,k_{\mathrm{int}}}(k_{\mathrm{int}}) = U(k_{\mathrm{int}}, [0, 3]), \tag{41}$$
$$p_{0,r}(r) = U(r, [0, 50 \times 10^6]), \tag{42}$$

Fig. 11 shows the posterior distribution of the model parameters $\beta_0, \gamma, k_{\mathrm{int}}$ obtained using rejection sampling with 10^5 samples. It is observed that the inferred parameters are correlated with marginal PDFs that differ significantly from the prior.

We implement CEBU with the GM and vMFNM models as the parametric density. The target coefficient of variation of the weights is set to $\delta_{target} = 1$ and $K = 3$ mixture terms are employed. The performance of the method is assessed by comparing the posterior means of β_0, γ and k_{int}, and the log-marginal likelihood, $\log Z_{\boldsymbol{y}}$, with the ones obtained from the reference solution (based on 10^5 samples drawn through rejection sampling), aBUS and SMC. The simulation results for $N_S = 500, 2000$ and 8000 samples per level are reported in Tab. 2. Fig. 12 illustrates the marginal likelihood error and the required number of model evaluations for different values of N_S ranging between 500 and 8000. From Tab. 2, it is observed that the estimates of the posterior means and marginal likelihood obtained from CEBU improve with increase in the number of samples per level, however the convergence is significantly faster for the GM density. The estimates from CEBU-GM with
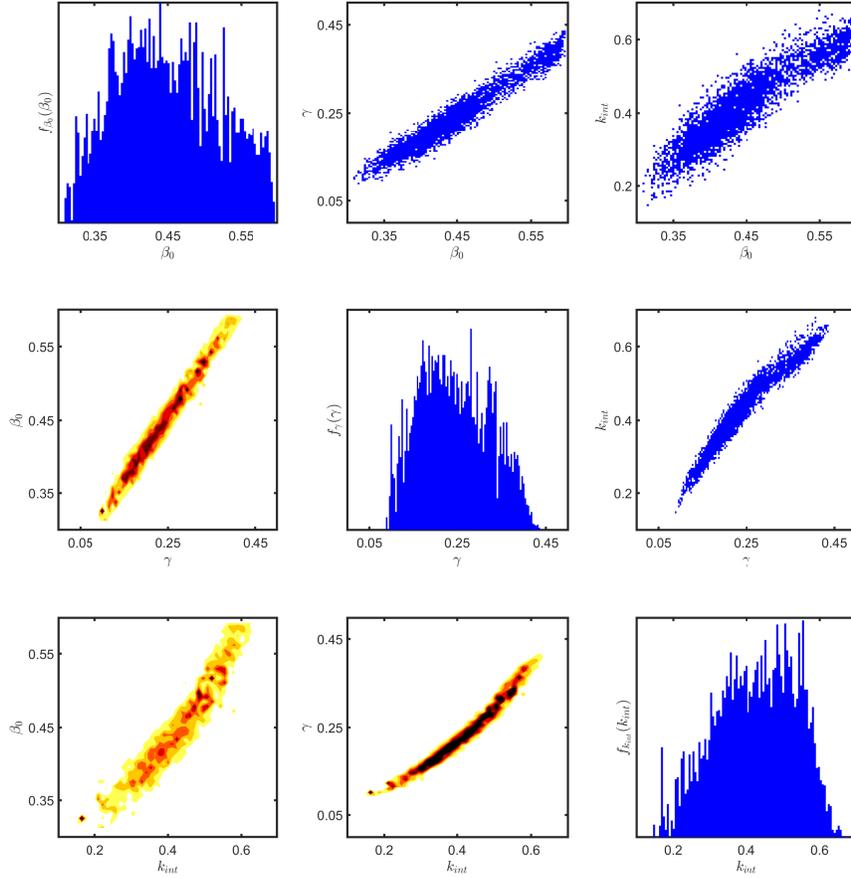
Figure 11: Posterior PDF of the basic infection rate β_0, removal rate γ and reduction factor k_{int}: samples drawn from the posterior distribution (upper triangle), marginal distribution of the individual parameters (diagonal), heat map of the sample approximation of the posterior distribution (lower triangle).

$N_S = 2000$ agree sufficiently well with the reference solution, whereas for CEBU-vMFNM a larger N_S is required for similar compliance. Furthermore, the GM density outperforms the vMFNM density in terms of efficiency since it achieves a lower c.o.v. of the estimates with a significantly smaller number

of model evaluations. As with example 5.2, the poor performance of CEBU-vMFNM for this example is attributed to the inability of the vMFNM density to describe accurately highly correlated posterior PDFs. From Tab. 2, it is observed that the estimates from aBUS and SMC compare well with the reference solution and both approaches show broadly similar performance. In comparison with CEBU-GM, we observe that these methods require a smaller number of model evaluations, however the corresponding c.o.v. of the estimates is about one order of magnitude higher. The number of model evaluations in CEBU-GM is higher than in SMC due to the limited number of mixture terms used, which cause CEBU-GM to require a larger number of iterations for convergence to the posterior PDF. The relative performance of the methods can be further assessed from Fig. 12, where it is seen that the CE method with the GM density gives a much narrower confidence interval for the marginal likelihood error than other approaches. Fig. 12b shows that aBUS requires the least number of model evaluations followed by SMC and CEBU-GM. CEBU-vMFNM requires a much larger number of model evaluations than the other approaches.

| (a) | (b) |
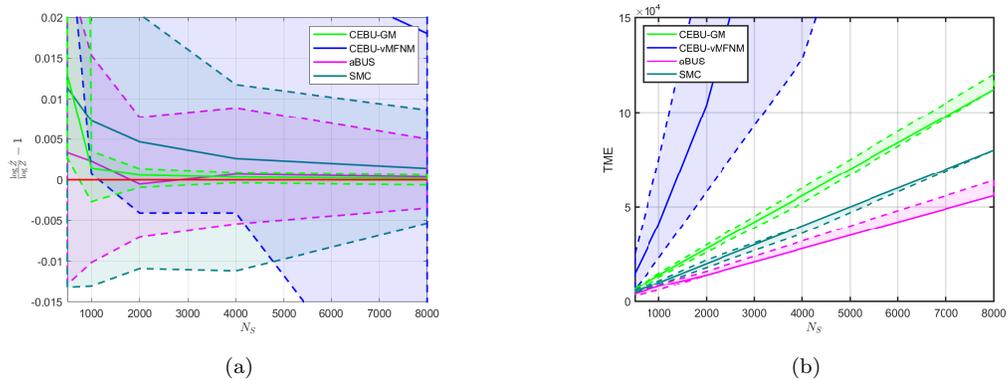
Figure 12: a: Log-marginal likelihood error for different number of samples per level, N_S, in example 5.3. - b: Total model evaluations for different number of samples per level, N_S, in example 5.3. The continuous lines show the medians and the shaded areas show the 95% credible intervals obtained from 100 independent simulation runs.

Finally, in Fig. 13 we show the 95% credible interval and the median of the posterior predicted number of infected people per day, as obtained from CEBU-GM, against the data. An acceptable match between the predictions from CEBU-GM and the real data is observed, with the predicted mean intervention day, $\left(E\left[t_{\text{int}} - \dfrac{\tau_{\text{int}}}{2} \right] \right)$, 2021-03-24, being close to the official first
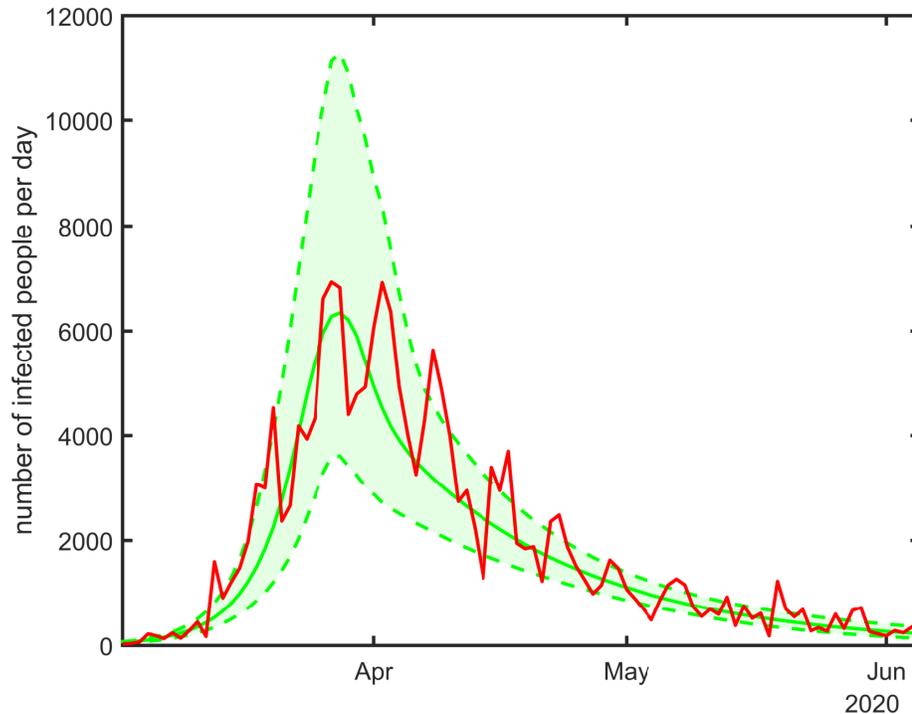
Figure 13: Median and 95%-credible interval of the number of infected people per day predicted by the posterior samples of CEBU-GM (K=3) against the data (red line) of corona example 5.3

day of lockdown in Germany, which was 2021-03-23. It is noted that we are aware of the fact that the intervention in our epidemiologic model might be related to high order effects, which happened before the specific date, since the data is subjected to some temporal delay of report.

6. Conclusion

We propose an adaptive importance sampling (IS) method for Bayesian updating of model parameters, in which an effective IS density is determined using the principle of cross entropy (CE) minimization. A tempering of the

Table 2: Estimates of the posterior mean of the model parameters, the log-marginal likelihood and the total model evaluations for all investigated methods, with different choices of the number of samples per level, N_s, in example 5.3. The c.o.v./std of the estimates is obtained from 100 independent simulation runs.

Method	N_S	$\hat{\mu}_\beta$ (c.o.v.)	$\hat{\mu}_\gamma$ (c.o.v.)	$\hat{\mu}_{k_{int}}$ (c.o.v.)	$\log \hat{Z}$ (std.)	TME (c.o.v.)
CEBU-GM	500	4.01e-01 (8.58e-02)	2.08e-01 (1.71e-01)	3.97e-01 (1.91e-01)	-122.17e+00 (10.75e+00)	6315 (1.16e-01)
	2000	4.42e-01 (1.55e-02)	2.36e-01 (3.36e-02)	4.16e-01 (2.64e-02)	-116.04e+00 (7.10e-02)	27900 (3.58e-02)
	8000	4.46e-01 (1.24e-02)	2.40e-01 (2.38e-02)	4.21e-01 (1.71e-02)	-115.99e+00 (4.91e-02)	115520 (3.98e-02)
CEBU-vMFNM	500	4.07e-01 (1.53e-01)	2.07e-01 (3.15e-01)	4.00e-01 (3.96e-01)	-147.69e+00 (42.00e+00)	15330 (3.30e-01)
	2000	4.25e-01 (1.30e-01)	2.14e-01 (2.79e-01)	3.90e-01 (2.38e-01)	-132.05e+00 (22.45e+00)	106860 (2.98e-01)
	8000	4.39e-01 (9.51e-02)	2.32e-01 (1.90e-01)	4.16e-01 (1.48e-01)	-122.70e+00 (16.81e+00)	698880 (1.72e-01)
aBUS	500	4.28e-01 (1.21e-01)	2.20e-01 (2.69e-01)	3.87e-01 (2.21e-01)	-116.47e+00 (1.15e+00)	3785 (8.66e-02)
	2000	4.47e-01 (6.46e-02)	2.41e-01 (1.33e-01)	4.19e-01 (1.04e-01)	-116.00e+00 (4.52e-01)	14720 (7.36e-02)
	8000	4.48e-01 (3.28e-02)	2.43e-01 (6.91e-02)	4.24e-01 (5.44e-02)	-116.01e+00 (2.37e-01)	59600 (6.98e-02)
SMC	500	4.34e-01 (1.40e-01)	2.27e-01 (3.03e-01)	3.98e-01 (2.40e-01)	-117.71e+00 (1.84e+00)	5150 (7.10e-02)
	2000	4.44e-01 (8.81e-02)	2.38e-01 (1.84e-01)	4.14e-01 (1.44e-01)	-116.53e+00 (8.26e-01)	20040 (3.75e-02)
	8000	4.43e-01 (5.37e-02)	2.37e-01 (1.19e-01)	4.16e-01 (9.59e-02)	-116.13e+00 (4.71e-01)	79840 (1.41e-02)
Reference	-	4.49e-01	2.43e-01	4.23e-01	-115.98e+00	40740592

likelihood function is introduced to efficiently solve the CE optimization problem. The implementation of the CE method requires a suitable parametric IS density. This parametric density is nudged towards the optimal IS density by an adaptive selection of the tempering parameter based on an effective sample size criterion. The final IS density obtained from the CE method is also a good approximation of the posterior probability density function (PDF) of the uncertain model parameters, and is used to generate posterior samples by a resampling scheme.

We implement the CE method using the Gaussian mixture (GM) and von Mises-Fisher-Nakagami mixture (vMFNM) models as the parametric family. Their performance is investigated in the context of three numerical examples. In terms of the confidence per run, the proposed method is found to be more efficient than other existing approaches, such as adaptive Bayesian updating with subset simulation and sequential Monte Carlo methods. In particular, the CE method for Bayesian updating using the GM distribution showed remarkable performance. The vMFNM model is highly efficient for high dimensional posterior PDFs with independent or weakly dependent marginal distributions. However, its performance for problems where the posterior PDF has highly correlated marginals is poor.

As future research, one could explore more flexible parametric importance sampling densities, like the student-t mixture model or an adapted version of the vMFNM distribution, in the context of the proposed method. Furthermore, a density based clustering algorithm considering the importance weights could be tested for the expectation-maximization algorithm for es-

timating the parameters of the IS density. Due to the inherently sequential structure of the proposed method, further applications of the method to Bayesian filtering problems could be explored, wherein the final IS density obtained from the CE method could be used as an approximate prior for the next step. Finally, we note that the proposed method allows full parallelization of the likelihood function evaluations. This provides a scope for further enhancing the efficiency of the method.

Acknowledgements

References

[1] W. R. Gilks, S. Richardson, D. Spiegelhalter, Markov chain Monte Carlo in practice, CRC press, 1995.

[2] R. M. Neal, Annealed importance sampling, Statistics and Computing 11 (2001) 125–139.

[3] N. Chopin, A sequential particle filter method for static models, Biometrika 89 (2002) 539–552.

[4] J. L. Beck, S.-K. Au, Bayesian updating of structural models and reliability using markov chain monte carlo simulation, Journal of Engineering Mechanics 128 (2002) 380–391.

[5] P. Del Moral, A. Doucet, A. Jasra, Sequential Monte Carlo samplers, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 411–436.

[6] S. L. Cotter, G. O. Roberts, A. M. Stuart, D. White, Mcmc methods for functions: modifying old algorithms to make them faster, Statistical Science 28 (2013) 424–446.

[7] I. Papaioannou, W. Betz, K. Zwirglmaier, D. Straub, Mcmc algorithms for subset simulation, Probabilistic Engineering Mechanics 41 (2015) 89–103.

[8] Z. Hu, Z. Yao, J. Li, On an adaptive preconditioned crank–nicolson mcmc algorithm for infinite dimensional bayesian inference, Journal of Computational Physics 332 (2017) 492–503.

[9] A. Beskos, A. Jasra, K. Law, Y. Marzouk, Y. Zhou, Multilevel sequential monte carlo with dimension-independent likelihood-informed proposals, SIAM/ASA Journal on Uncertainty Quantification 6 (2018) 762–786.

[10] B. Carrera, C. M. Mok, I. Papaioannou, Efficient estimation of hydraulic conductivity heterogeneity with non-redundant measurement information, GEM-International Journal on Geomathematics 11 (2020) 1–29.

[11] J. Ching, Y.-C. Chen, Transitional markov chain monte carlo method for bayesian model updating, model class selection, and model averaging, Journal of Engineering Mechanics 133 (2007) 816–832.

[12] A. Jasra, D. A. Stephens, A. Doucet, T. Tsagaris, Inference for lévy-driven stochastic volatility models via adaptive sequential monte carlo, Scandinavian Journal of Statistics 38 (2011) 1–22.

[13] W. Betz, I. Papaioannou, D. Straub, Transitional markov chain monte carlo: observations and improvements, Journal of Engineering Mechanics 142 (2016) 04016016.

[14] D. Straub, I. Papaioannou, Bayesian updating with structural reliability methods, Journal of Engineering Mechanics 141 (2015) 04014134.

[15] F. DiazDelaO, A. Garbuno-Inigo, S. Au, I. Yoshida, Bayesian updating and model class selection with subset simulation, Computer Methods in Applied Mechanics and Engineering 317 (2017) 1102–1121.

[16] W. Betz, I. Papaioannou, J. L. Beck, D. Straub, Bayesian inference with subset simulation: strategies and improvements, Computer Methods in Applied Mechanics and Engineering 331 (2018) 72–93.

[17] A. B. Owen, Monte Carlo theory, methods and examples, 2013.

[18] O. Cappé, R. Douc, A. Guillin, J.-M. Marin, C. P. Robert, Adaptive importance sampling in general mixture classes, Statistics and Computing 18 (2008) 447–459.

[19] M. Morzfeld, M. S. Day, R. W. Grout, G. S. Heng Pau, S. A. Finsterle, J. B. Bell, Iterative importance sampling algorithms for parameter estimation, SIAM Journal on Scientific Computing 40 (2018) B329–B352.

[20] J. L. Beck, L. S. Katafygiotis, Updating models and their uncertainties. i: Bayesian statistical framework, Journal of Engineering Mechanics 124 (1998) 455–461.

[21] C. Papadimitriou, J. L. Beck, L. S. Katafygiotis, Updating robust reliability using structural test data, Probabilistic Engineering Mechanics 16 (2001) 103–113.

[22] R. Y. Rubinstein, D. P. Kroese, Simulation and the Monte Carlo method, volume 10, John Wiley & Sons, 2016.

[23] R. Y. Rubinstein, Optimization of computer simulation models with rare events, European Journal of Operational Research 99 (1997) 89–112.

[24] K.-P. Hui, N. Bean, M. Kraetzl, D. P. Kroese, The cross-entropy method for network reliability estimation, Annals of Operations Research 134 (2005) 101.

[25] Z. I. Botev, D. P. Kroese, T. Taimre, Generalized cross-entropy methods with applications to rare-event simulation and optimization, Simulation 83 (2007) 785–806.

[26] N. Kurtz, J. Song, Cross-entropy-based adaptive importance sampling using Gaussian mixture, Structural Safety 42 (2013) 35–44.

[27] S. Geyer, I. Papaioannou, D. Straub, Cross entropy-based importance sampling using gaussian densities revisited, Structural Safety 76 (2019) 15–27.

[28] I. Papaioannou, S. Geyer, D. Straub, Improved cross entropy-based importance sampling with a flexible mixture model, Reliability Engineering & System Safety 191 (2019) 106564.

[29] O. Kanjilal, I. Papaioannou, D. Straub, Cross entropy-based importance sampling for first-passage probability estimation of randomly excited linear structures with parameter uncertainty, Structural Safety 91 (2021) 102090.

[30] O. Kanjilal, I. Papaioannou, D. Straub, Series system reliability of uncertain linear structures under gaussian excitation by cross entropy-based importance sampling, ASCE Journal of Engineering Mechanics 148 (1) (2022) 04021136.

[31] F. Uribe, I. Papaioannou, Y. M. Marzouk, D. Straub, Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation, SIAM/ASA Journal on Uncertainty Quantification 9 (2021) 818–847.

[32] R. Rubinstein, The cross-entropy method for combinatorial and continuous optimization, Methodology and Computing in Applied Probability 1 (1999) 127–190.

[33] J. C. Chan, E. Eisenstat, Marginal likelihood estimation with the cross-entropy method, Econometric Reviews 34 (2015) 256–285.

[34] J. C. Chan, D. P. Kroese, Improved cross-entropy method for estimation, Statistics and Computing 22 (2012) 1031–1040.

[35] M. Dashti, A. M. Stuart, The Bayesian approach to inverse problems, in: R. Ghanem, D. Higdon, H. Owhadi (Eds.), Handbook of Uncertainty Quantification, Springer, Cham, 2017, pp. 311–428.

[36] J. L. Beck, K.-V. Yuen, Model selection using response measurements: Bayesian probabilistic approach, Journal of Engineering Mechanics 130 (2004) 192–203.

[37] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

[38] A. Beskos, A. Jasra, N. Kantas, A. Thiery, et al., On the convergence of adaptive sequential monte carlo methods, Annals of Applied Probability 26 (2016) 1111–1146.

[39] J. Latz, I. Papaioannou, E. Ullmann, Multilevel sequential monte carlo for bayesian inverse problems, Journal of Computational Physics 368 (2018) 154–178.

[40] D. Randal, O. Cappé, E. Moulines, Comparison of resampling schemes for particle filtering, in: Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005, pp. 64–69.

[41] Y. Li, W. Wang, K. Deng, J. S. Liu, Stratification and optimal re-sampling for sequential monte carlo, Biometrika (Forthcoming) (????). doi:10.1093/biomet/asab004.

[42] V. Elvira, L. Martino, D. Luengo, M. F. Bugallo, Novel weighting and resampling schemes in population monte carlo, in: XXVIème Colloque GRETSI, 2017.

[43] W. K. Hastings, Monte carlo sampling methods using markov chains and their applications, Biometrika 1 (1970) 97–109.

[44] P.-L. Liu, A. Der Kiureghian, Multivariate distribution models with pre-scribed marginals and covariances, Probabilistic Engineering Mechanics 1 (1986) 105–112.

[45] M. Hohenbichler, R. Rackwitz, Non-normal dependent vectors in struc-tural safety, ASCE Journal of the Engineering Mechanics Division 107 (1981) 1227–1238.

[46] L. S. Katafygiotis, K. M. Zuev, Geometric insight into the challenges of solving high-dimensional reliability problems, Probabilistic Engineering Mechanics 23 (2008) 208–218.

[47] G. Biau, D. M. Mason, High-dimensional p-norms, in: M. Hallin, D. M. Mason, D. Pfeifer, J. G. Steinebach (Eds.), Mathematical Statistics and Limit Theorems, Springer International Publishing, Cham, 2015, pp. 21–40.

[48] A. F. M. Smith, A. E. Gelfand, Bayesian statistics without tears: A sampling-resampling perspective, The American Statistician 46 (1992) 84.

[49] W. Betz, J. L. Beck, I. Papaioannou, D. Straub, Bayesian inference with reliability methods without knowing the maximum of the likelihood function, Probabilistic Engineering Mechanics 53 (2018) 14–22.

[50] Dimitris G. Giovanis, Iason Papaioannou, Daniel Straub, Vissarion Pa-padopoulos, Bayesian updating with subset simulation using artificial neural networks, Computer Methods in Applied Mechanics and Engi-neering 319 (2017) 124–145.

[51] A. K. Chopra, Dynamics of structures: theory and application to earthquake engineering, 5th Edition, Pearson Education, 2016.

[52] P. Karnakov, G. Arampatzis, I. Kičić, F. Wermelinger, D. Wälchli, C. Papadimitriou, P. Koumoutsakos, Data-driven inference of the reproduction number for covid-19 before and after interventions for 51 european countries, Swiss medical weekly 150 (2020) w20313.

[53] H. W. Hethcote, The mathematics of infectious diseases, SIAM review 42 (2000) 599–653.

[54] CSSE at Johns Hopkins University, Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university, 26.02.2021.

Appendix A. Evaluation of the marginal likelihood in the standard normal space

Using the relation in Eq. 2 and applying the iso-probabilistic transformation between the physical space and the standard normal space, one obtains

$$
\begin{aligned}
Z_{\boldsymbol{y}} &= \int L_{\boldsymbol{y}}(\boldsymbol{\theta}) p_0(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int L_{\boldsymbol{y}}\left(T_{ISO}^{-1}(\boldsymbol{u})\right) p_0\left(T_{ISO}^{-1}(\boldsymbol{u})\right) |\det(\boldsymbol{J}_{T_{ISO}}^{-1})| d\boldsymbol{u} \qquad \text{(A.1)} \\
&= \int L_{\boldsymbol{y}}\left(T_{ISO}^{-1}(\boldsymbol{u})\right) \phi(\boldsymbol{u}) d\boldsymbol{u}
\end{aligned}
$$

as the marginal likelihood, where $\boldsymbol{J}_{T_{ISO}}$ denotes the Jacobian of the transformation T_{ISO}. On applying the IS identity to Eq. A.1 with $h(\boldsymbol{u}, \hat{\boldsymbol{\nu}}_T)$ as the IS density, the following relationship is obtained:

$$
Z_{\boldsymbol{y}} = \int W(\boldsymbol{u}, \hat{\boldsymbol{\nu}}_T) h(\boldsymbol{u}, \hat{\boldsymbol{\nu}}_T) d\boldsymbol{u}, \qquad \text{(A.2)}
$$

where $W(\boldsymbol{u}, \hat{\boldsymbol{\nu}}_T) = \frac{L_{\boldsymbol{y}}(T_{ISO}^{-1}(\boldsymbol{u}))\phi_0(\boldsymbol{u})}{h(\boldsymbol{u},\hat{\boldsymbol{\nu}}_T)}$. Here $\hat{\boldsymbol{\nu}}_T$ denotes the parameter vector obtained in the final tempering step of the CE method. $Z_{\boldsymbol{y}}$ can then be estimated in accordance with Eq. 19 using samples from $h(\boldsymbol{u}, \hat{\boldsymbol{\nu}}_T)$.