Transitional Markov Chain Monte Carlo: Observations and Improvements

Wolfgang Betz*, Iason Papaioannou, Daniel Straub

Engineering Risk Analysis Group, Technische Universität München, 80333 München, Germany

Abstract

The Transitional Markov Chain Monte Carlo (TMCMC) method is a widely used method for Bayesian updating and Bayesian model class selection. The method is based on successively sampling from a sequence of distributions that gradually approach the posterior target distribution. The samples of the intermediate distributions are used to obtain an estimate of the evidence, which is needed in the context of Bayesian model class selection. We discuss the properties of the TMCMC method and propose the following three modifications to the TMCMC method: (1) The sample weights should be adjusted after each MCMC step. (2) A burn-in period in the MCMC sampling step can improve the posterior approximation. (3) The scale of the proposal distribution of the MCMC algorithm can be selected adaptively to achieve a near-optimal acceptance rate. We compare the performance of the proposed modifications with the original TMCMC method by means of three example problems. The proposed modifications reduce the bias in the estimate of the evidence, and improve the convergence behavior of posterior estimates.

Keywords: Bayesian updating, evidence, transitional Markov chain Monte Carlo, MCMC, bias, burn-in

Introduction

Bayesian analysis provides a consistent framework to reduce uncertainties in existing models through new information. The uncertainties in the model are expressed by means of input parameters that are treated as random. New information from observations or measurements is then used to update our prior belief about the parameters of the model to a posterior belief. If the posterior model cannot be derived analytically, samples from the posterior have to be generated numerically.

Sampling from the posterior distribution is not a straightforward task because the posterior is typically only know pointwise in terms of the outcome of a numerical model. Usually, sampling from the posterior is performed through methods that are based on Markov Chain Monte Carlo (MCMC) sampling [12, 11]. Application of MCMC approaches requires awareness of the following issues: First, all methods that employ MCMC sampling produce dependent samples. Thus, the efficiency of statistical estimators is reduced, compared to crude Monte Carlo methods that produce independent samples. Second, a Markov chain needs to be initialized with a seed. If this seed is not already a sample from the posterior distribution, the chain needs a certain number of steps until the produced samples approximately follow the posterior distribution. The initial phase until the Markov chain produces samples from the posterior is called *burn-in* period. In practice, the length of the burn-in period is difficult to assess. Third, many MCMC techniques become inefficient when the number of uncertain parameters in the Bayesian analysis becomes large, because it is difficult to set-up a transition density that works well in high dimensions.

The Transitional Markov Chain Monte Carlo (TMCMC) method, proposed by Ching and Chen (2007), belongs to the class of sequential particle filter methods [9] and is based on MCMC sampling. The method tries to overcome the aforementioned issues of MCMC by gradually pushing the samples from the prior to the posterior distribution. The method has become popular in both research and practice: recent contributions include [25], [15], [17], [13], [2]. In addition to the posterior samples generated by TMCMC, the method returns an estimate of the evidence of the Bayesian model class, which is needed for Bayesian model class selection and Bayesian model averaging [16, 24].

In this contribution, we discuss properties of the TMCMC method, and on this basis we identify possible improvements. In particular, we observe that the TMCMC method tends to produce estimates of the evidence that contain a considerable bias. Three potential modifications to the TMCMC method are proposed: (1) Adjusting the sample weights after each MCMC step tends to improve the performance of the method and reduces the bias in the estimated evidence. (2) A burn-in period in the MCMC sampling step can improve the posterior approximation. (3) The scale of the proposal distribution can be adjusted adaptively such that the MCMC algorithm maintains a specified near-optimal acceptance rate.

The outline of the paper is as follows: First, we introduce Bayesian model updating in general and the TMCMC method in particular. Thereafter, we make a number of observations on the performance of the TMCMC method and propose modifications to improve it. Finally, the effects of the proposed modifications are showcased through a comparison with the original

*Corresponding author

Email addresses: wolfgang.betz@tum.de (Wolfgang Betz), iason.papaioannou@tum.de (Iason Papaioannou), straub@tum.de (Daniel Straub)

TMCMC method by means of three numerical examples.

Bayesian model updating

In Bayesian inference, the learning process is formalized through Bayes' theorem:

$$p(\theta|\mathbf{d}) = c_{\mathrm{E}}{}^{-1} \cdot L(\theta|\mathbf{d}) \cdot p(\theta) \tag{1}$$

where θ is the vector of M model parameters that are considered uncertain in the analysis, and $\mathbf{d}$ denotes the new information that becomes available in form of measurements or observations. The probability density function (PDF) $p(\theta)$ represents our prior belief about the distribution of θ, the likelihood $L(\theta|\mathbf{d})$ expresses the plausibility of observing $\mathbf{d}$ given a certain θ, $p(\theta|\mathbf{d})$ is the resulting posterior distribution, and c_{E} is a scaling constant defined as:

$$c_{\mathrm{E}} = \int_{\mathbf{d}} L(\theta|\mathbf{d}) p(\theta)\, \mathrm{d}\theta \tag{2}$$

The constant c_{E} is a measure for the plausibility of the investigated model class. In the Bayesian community c_{E} is usually referred to as the evidence of the model class [4]. In case of multiple model classes, the evidence allows evaluating the posterior plausibilities of the individual model classes. The plausibilities are required for Bayesian model class selection and Bayesian model averaging [16, 24]. Therefore, it is advantageous if an estimate of c_{E} is obtained as a by-product of the method that performs the Bayesian inference.

TMCMC method

The principle behind TMCMC

The TMCMC method [8] belongs to the class of sequential particle filter methods [9]. The method starts with independent samples from the prior distribution. In subsequent steps, the sampling distribution is gradually transformed such that it approaches the posterior distribution. For this purpose, (Eq. 1) is modified to:

$$p_j(\theta) \propto p(\theta) \cdot L(\theta|\mathbf{d})^{q_j} \tag{3}$$

where $j = 0, \ldots, m$ denotes the level, and the $q_j \in [0, 1]$ are chosen such that $q_0 = 0 < q_1 < \cdots < q_m = 1$. Consequently, for $j = 0$, $p_0(\theta)$ is equal to the prior distribution $p(\theta)$; and for $j = m$, $p_m(\mathbf{x})$ matches the posterior distribution $p(\theta|\mathbf{d})$.

The principle behind the TMCMC method is to gradually push the samples from the prior distribution to the posterior distribution. The speed of this gradual transition is controlled by the coefficients q_j. Ching and Chen (2007) proposed to select q_{j+1} based on q_j such that the coefficient of variation of $L(\theta|\mathbf{d})^{q_{j+1}-q_j}$ approximately equals v_{t}, where $v_{\mathrm{t}} = 100\%$ was suggested. The value of q_{j+1} can then be determined based on the samples of the previous level as:

$$q_{j+1} = \arg\min_{q} \left(\left| \mathrm{CV}_j(q) - v_{\mathrm{t}} \right| \right) \tag{4}$$

where $\mathrm{CV}_j(q)$ with $q \in (q_j, 1]$ is the sample coefficient of variation of the set $\left\{ L(\theta_{(j,k)}|\mathbf{d})^{q-q_j} \right\}_{k=1}^{N_{\mathrm{s}}}$, N_{s} is the number of samples generated at each level, $\theta_{(j,k)}$ denotes the kth sample at level j, and $L(\theta_{(j,k)}|\mathbf{d})$ is the likelihood value that is associated with $\theta_{(j,k)}$.

The evidence of the stochastic model class c_{E} (Eq. 2) can be rewritten as follows:

$$c_{\mathrm{E}} = \int_{\mathbf{x}} p(\theta) \cdot \prod_{j=1}^{m} L(\theta|\mathbf{d})^{q_j - q_{j-1}}\, \mathrm{d}\theta \tag{5}$$

$$= \prod_{j=1}^{m} \int_{\theta} L(\theta|\mathbf{d})^{q_j - q_{j-1}} \cdot p_{j-1}(\theta)\, \mathrm{d}\theta \tag{6}$$

$$= \prod_{j=1}^{m} \mathrm{E}_{p_{j-1}(\theta)} \left[L(\theta|\mathbf{d})^{q_j - q_{j-1}} \right] \tag{7}$$

where $\mathrm{E}_{p_{j-1}(\theta)}[\cdot]$ denotes the expectation with respect to distribution p_{j-1}, which can be estimated based on the generated samples:

$$\mathrm{E}_{p_j(\theta)} \left[L(\theta|\mathbf{d})^{q_{j+1}-q_j} \right] \approx \frac{1}{N_{\mathrm{s}}} \sum_{k=1}^{N_{\mathrm{s}}} L(\theta_{(j,k)}|\mathbf{d})^{q_{j+1}-q_j} \tag{8}$$

The original TMCMC method proposed by Ching and Chen (2007)

The TMCMC algorithm can be summarized as follows: For the initial $j = 0$, all N_{s} samples $\theta_{(0,1)}, \ldots, \theta_{(0,N_{\mathrm{s}})}$ are drawn from the prior distribution, and j is set to *one* thereafter. For all $j > 0$, Ching and Chen (2007) propose the following scheme:

1. Find q_j through solving (Eq. 4). If $q_j > 1$, then set $q_j = 1$.
2. For all samples $k = 1, \ldots, N_{\mathrm{s}}$ compute a weighting coefficient $w_{(j,k)}$:

$$w_{(j,k)} = \left(L(\theta_{(j-1,k)}|\mathbf{d}) \right)^{q_j - q_{j-1}} \tag{9}$$

3. Compute the mean of the weighting coefficients:

$$S_j = \frac{1}{N_{\mathrm{s}}} \sum_{k=1}^{N_{\mathrm{s}}} w_{(j,k)} \tag{10}$$

4. Compute the covariance matrix of the Gaussian proposal distribution:

$$\mathbf{\Sigma}_j = \beta^2 \cdot \sum_{k=1}^{N_{\mathrm{s}}} \left[\frac{w_{(j,k)}}{S_j \cdot N_{\mathrm{s}}} \cdot \left(\theta_{(j-1,k)} - \overline{\theta}_j \right) \cdot \left(\theta_{(j-1,k)} - \overline{\theta}_j \right)^{\mathrm{T}} \right] \tag{11}$$

with

$$\overline{\theta}_j = \frac{\sum_{l=1}^{N_{\mathrm{s}}} w_{(j,l)} \cdot \theta_{(j-1,l)}}{\sum_{l=1}^{N_{\mathrm{s}}} w_{(j,l)}} \tag{12}$$

The coefficient β scales the proposal distribution. Ching and Chen (2007) suggest to set $\beta = 0.2$.

5. For each l in $\{1, \ldots, N_{\mathrm{s}}\}$ set: $\theta^{\mathrm{c}}_{(j,l)} = \theta_{(j-1,l)}$. Thereafter, for $k = 1, \ldots, N_{\mathrm{s}}$ do:

- Select index l from the set $\{1, \ldots, N_s\}$ at random, where each l is assigned probability $\frac{w_{(j,l)}}{\sum_{n=1}^{N_s} w_{(j,n)}}$.
- Propose a new sample: draw θ^c from a Normal distribution that is centered at $\theta^c_{(j,l)}$ and has covariance matrix $\mathbf{\Sigma}_j$.
- Generate a sample r from a uniform distribution on $[0, 1]$.
- If $r \leq \frac{p_j(\theta^c)}{p_j(\theta^c_{(j,l)})}$ then set $\theta^c_{(j,l)} = \theta^c$, otherwise do nothing.
- Set $\theta_{(j,k)} = \theta^c_{(j,l)}$.

6. If $q_j = 1$ then stop the iteration, otherwise set $j = j + 1$ and continue with 1.

An estimate for the evidence of the assumed model class that is based on (Eq. 7) is:

$$\hat{c}_E = \prod_{j=1}^{m} S_j \tag{13}$$

with S_j defined according to (Eq. 10).

Observation 1: sample weights

At each level j in the TMCMC method, samples $\{\theta_{(j,1)}, \ldots, \theta_{(j,N_s)}\}$ are generated that (approximately) follow distribution p_j based on samples $\{\theta_{(j-1,1)}, \ldots, \theta_{(j-1,N_s)}\}$ that follow distribution p_{j-1}. To each sample a weight is attached according to (Eq. 9). Instead of performing only a simple resampling step based on the weights, the weighted sampling is combined with a MCMC step: In an iterative process, one randomly picks a sample according to its weight, uses the selected sample as seed to perform a MCMC step with stationary distribution p_j, and replaces the randomly picked sample with the sample that the MCMC step produced. The principle is that the samples that the MCMC step produced already follow (asymptotically) the target distribution p_j, because the seeds of the MCMC step are picked according to their importance weights.

If the proposed sample is accepted, the chain moves on. Therefore, the absolute weight of the current chain should change. This is not taken into account by the original TMCMC method. Asymptotically, the samples of the individual Markov chains that initially follow distribution p_{j-1} approach the target distribution p_j, and the weights of all chains should asymptotically equalize. However, in practice the intermediate distributions of the transition are difficult to obtain. Therefore, the original TMCMC method assumes that for the finite number of samples that are drawn from a single chain, no transition takes place. Consequently, the absolute weight of a Markov chain can be updated after each accepted sample by means of (Eq. 9). To consider the transition to the target distribution when updating the weights remains an area of future research.

Proposed modification (1):. In the original TMCMC method, the weights $w_{(j,k)}$ attached to each Markov chain are computed at the beginning of each TMCMC level (see (Eq. 9)) and then kept constant. We propose to adapt the weight of a Markov chain each time the chain moves on. In order to do so, step (5) in the TMCMC algorithm given above needs to be modified as follows:

5. For each l in $\{1, \ldots, N_s\}$ set: $\theta^c_{(j,l)} = \theta_{(j-1,l)}$. Thereafter, for $k = 1, \ldots, N_s$ do:
   - ...
   - Set $\theta_{(j,k)} = \theta^c_{(j,l)}$.
   - Set $w_{(j,l)} = \left(L(\theta^c_{(j,l)}|\mathbf{d})\right)^{q_j - q_{j-1}}$

Note that S_j and $\mathbf{\Sigma}_j$ are computed only once at the beginning of each level.

As we will demonstrate by means of numerical examples, this modification considerably reduces the average bias in the estimate of the evidence. The statistics of the posterior samples is improved marginally by the modification.

Observation 2: burn-in

In the TMCMC method one uses samples from distribution p_{j-1} to generate samples that asymptotically follow distribution p_j. The sampling is based on MCMC, where weighted samples of distribution p_{j-1} are taken as seeds. The weighted samples follow distribution p_j only asymptotically. Therefore, the MCMC sampling performed in the TMCMC method does not possess the property of *perfect sampling* [7], where *perfect sampling* implies that the initial distribution of the seeds equals the stationary distribution of the Markov chains [19].

For practical applications, this is usually not an issue, as we will demonstrate in numerical examples. Consequently, the TMCMC method does not usually require a burn-in period. However, one should be cautious if TMCMC is used with only a small number of samples per level: In this case a burn-in period might actually be required for the Markov chains to converge.

Proposed modification (2):. It is straight-forward to introduce a burn-in period of length N_b to the MCMC sampling of each TMCMC level. Again, only step (5) in the TMCMC algorithm has to be modified:

5. For each l in $\{1, \ldots, N_s\}$ set: $\theta^c_{(j,l)} = \theta_{(j-1,l)}$. Thereafter, for $k = 1, \ldots, (N_s + N_b)$ do:
   - ...
   - If $r \leq \frac{L(\theta^c|\mathbf{d})}{L(\theta^c_{(j,l)}|\mathbf{d})}$ then set $\theta^c_{(j,l)} = \theta^c$, otherwise do nothing.
   - If $k > N_b$ then set $\theta_{(j,k-N_b)} = \theta^c_{(j,l)}$, otherwise do nothing.
   - Set $w_{(j,l)} = \left(L(\theta^c_{(j,l)}|\mathbf{d})\right)^{q_j - q_{j-1}}$

However, our numerical investigations demonstrate that N_b can usually be set to *zero*.

Observation 3: scaling of the proposal

The optimal value of the constant β that is used to scale the proposal distribution depends considerably on the problem at hand. On the one hand, a β that is selected too small leads to a proposal distribution that accepts many samples; however, subsequent samples in a Markov chain are close to each other. Thus, the correlation in the chain is large and the produced samples will not properly propagate into the relevant domain for reasonable N_s. On the other hand, a β that is selected too large leads to many rejected samples and, thus, also results in a large chain correlation. The larger the correlation in a Markov chain, the less efficient is the sampling procedure, because the effective number of independent samples in the chain is reduced and, thus, the variance of applied estimators is bound to increase.

Setting $\beta = 0.2$ as proposed in Ching and Chen (2007) works well for some problems, but is, in our experience, far from optimal for other problems.

Proposed modification (3):. In the original TMCMC algorithm, the proposal distribution is set-up in the space spanned by θ: The proposal is a multivariate Normal distribution that is centered at the current state of the Markov chain and whose variance is defined according to (Eq. 11).

Without loss of generality, we propose to represent the joint prior PDF of the uncertain parameter vector θ in terms of an underlying vector $\mathbf{u} \in \mathbb{R}^M$ of independent standard Normal random variables. If the components of θ are a-priori independent, then the transformation $u_i \rightarrow \theta_i$ of the ith component of the uncertain parameter vector is given as: $\theta_i = F^{-1}_{\theta_i}(\Phi(u_i))$, where $F^{-1}_{\theta_i}(\cdot)$ denotes the inverse CDF of the prior distribution of the ith component of θ, and $\Phi(\cdot)$ denotes the CDF of the standard Normal distribution. If the components of θ are dependent, the marginal transformation based on the Nataf model [10] or the Rosenblatt transformation [14] can be used.

The updating problem is then solved in terms of $\mathbf{u}$. Performing the Bayesian inference in an underlying standard Normal space has numerical advantages: (1) The uncertainty in u_i is normalized, whereas the uncertainty in θ_i is usually not. (2) The support of θ may be bounded, whereas the support of $\mathbf{u}$ is not bounded.

Based on this, the proposal is then set-up in the space spanned by $\mathbf{u}$: it is a multivariate Normal distribution that is centered at the current state of the Markov chain and whose covariance matrix is defined as the sample covariance in terms of $\mathbf{u}$ (and not in terms of θ). We propose to select the initial scaling factor β of the thus obtained proposal distribution as: $\beta = 2.4/\sqrt{M}$ based on [11, 1].

Additionally, the performance of the MCMC algorithm can be enhanced, by adjusting the scaling factor β adaptively during the simulation. Often the scaling factor β is tuned such that the average acceptance probability of the MCMC algorithm approaches a specified target acceptance-rate t_{acr} [1, 18]. We suggest to adaptively modify β such that the monitored average acceptance-rate approaches the following target acceptance-rate $t_{\mathrm{acr}} = 0.21/M + 0.23$, i.e., $t_{\mathrm{acr}} = 0.44$ for $M = 1$, $t_{\mathrm{acr}} = 0.27$ for $M = 5$ and $t_{\mathrm{acr}} = 0.23$ for large M. This rule is based on the findings published in [20, 21].

The algorithm to adaptively update β is as follows: At the initial sampling level set $\beta_{(\mathrm{old})} = 2.4/\sqrt{M}$; in all other levels use the last value of $\beta_{(\mathrm{old})}$ from the previous sampling level. At the beginning of each sampling level, set $N_{\mathrm{adapt}} = 1$. Perform N_{a} MCMC steps. Thereafter, evaluate the coefficient $c_{\mathrm{a}} = (p_{\mathrm{acr}} - t_{\mathrm{acr}}) / \sqrt{N_{\mathrm{adapt}}}$, where p_{acr} is the mean acceptance-rate of the last N_{a} MCMC steps (i.e., the number of accepted samples divided by N_{a}), and t_{acr} denotes the target acceptance-rate. Modify β based on the value of c_{a}: set $\beta_{(\mathrm{new})} = \beta_{(\mathrm{old})} \cdot \exp(c_{\mathrm{a}})$. Increase the value of N_{adapt} by one, set $\beta_{(\mathrm{old})} = \beta_{(\mathrm{new})}$, perform another N_{a} MCMC steps and evaluate c_{a} again. Repeat this until the required number of samples is generated. We suggest to set $N_{\mathrm{a}} = 100$.

Note: In principle, the algorithm to choose the scaling factor β adaptively works also if the problem is solved directly in the space spanned by θ. However, for the following reasons, we expect a reduced efficiency in this case: On the one hand, working in the underlying $\mathbf{u}$-space facilitates the choice of an initial scaling factor that leads to a robust behavior of the algorithm for a large variety of problems. Moreover, if the support of θ is bounded and a Normal proposal distribution is used, some proposed samples have to be rejected simply because the proposed sample is not within the support of θ. On the other hand, solving the updating problem in terms of $\mathbf{u}$ adds an additional layer of complexity.

Numerical Investigations

In the following we compare the proposed modifications of the TMCMC method with the original TMCMC method. We term the method that takes the first two proposed modifications into account *weight-adjusted TMCMC* (wTMCMC), and the method that takes the first and the third proposed modification into account *improved TMCMC* (iTMCMC). The burn-in length N_{b} is set to *zero* in iTMCMC.

The comparison is performed by means of three example problems for Bayesian updating: a dynamic system with two degrees of freedom, an analytical example with a bimodal likelihood function, and an example where the likelihood is expressed in terms of the sum of Normal random variables. Within a single updating run we generate 10^3 posterior samples, i.e. $N_{\mathrm{s}} = 10^3$. By repeatedly solving the updating problem several times, we evaluate the mean and standard deviation of the estimated evidence (denoted $a_{c_{\mathrm{E}}}$ and $s_{c_{\mathrm{E}}}$), and the posterior mean and standard deviation of estimate $\overline{g}$ (denoted $a_{\overline{g}}$ and $s_{\overline{g}}$), where $\overline{g}$ is defined as:

$$\mathrm{E}\left[g(\theta)|\mathbf{d}\right] \approx \frac{1}{N_{\mathrm{s}}} \sum_{k=1}^{N_{\mathrm{s}}} g(\theta_{(m,k)}) = \overline{g} \tag{14}$$

The scalar quantity $g(\theta)$ acts as a placeholder for the quantity of interest in a particular example; its meaning will be specified for each example separately. The estimated posterior mean and standard deviation of quantity g is denoted a_g and s_g, respectively.

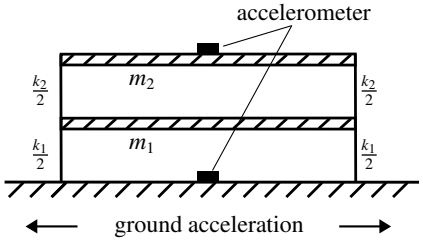


Figure 1: 2DOF system investigated as first example problem.

If the generated posterior samples were independent, one could write $\mathrm{Var}[\overline{g}] = s_{\overline{g}}^2 = \frac{1}{N_s} \cdot \sigma_g^2$, where σ_g is the *true* posterior standard deviation of scalar quantity $g(\Theta)$. However, the posterior samples are dependent, since they are generated by means of MCMC sampling. More specifically, the samples exhibit a positive dependence structure, which means $\mathrm{Var}[\overline{g}] = s_{\overline{g}}^2 > \frac{1}{N_s} \cdot \sigma_g^2$. An effective number of independent samples N_{eff} can be introduced such that $s_{\overline{g}}^2 = \frac{1}{N_{\mathrm{eff}}} \cdot \sigma_g^2$; i.e., on average, we could obtain the same $\mathrm{Var}[\overline{g}]$ (that we estimated using the N_s dependent posterior samples) if we have N_{eff} truly independent posterior samples at hand, with $N_{\mathrm{eff}} = \sigma_g^2 / s_{\overline{g}}^2$ and $N_{\mathrm{eff}} \leq N_s$. The total number of model calls required to generate N_s posterior samples is denoted as N_M, where $N_M = (m+1) \cdot N_s + m \cdot N_b$, with m being the number of intermediate levels, and N_b is the length of the burn-in period.

In order to investigate the bias in the estimate of the evidence, we compute

$$\mathrm{bias}_{c_E} = \frac{|a_{c_E} - c_E|}{c_E} \tag{15}$$

where a_{c_E} is the average of the evidence estimated by TMCMC, and c_E denotes the *true* evidence of the problem at hand. The overall quality of the estimate of the evidence is assessed by means of:

$$\kappa_{c_E} = \sqrt{(\mathrm{bias}_{c_E})^2 + \left(\frac{s_{c_E}}{a_{c_E}}\right)^2} \tag{16}$$

where s_{c_E} is the standard deviation of the estimated evidence. The measure κ_{c_E} considers the variability in the estimated evidence additional to the bias. κ_{c_E} can only be zero if the estimated evidence is always the true evidence. The bias in the estimated posterior mean and standard deviation is defined as $\mathrm{bias}_{a_g} = a_g / \mathrm{E}[g] - 1$ and $\mathrm{bias}_{s_g} = s_g / \sqrt{\mathrm{Var}[g]} - 1$, respectively.

Example problem 1: 2DOF dynamic problem

As a first example problem, we investigate the 2DOF system presented in the original TMCMC paper [8, Example 1, Case I]: The response of the two-story linear structure shown in Fig. 1 to a narrow banded ground acceleration is investigated. The ground acceleration and the acceleration of the roof are measured every 0.02 seconds for a period of one second. We assume that perfect measurements of the ground acceleration are available, however, the measurements of the roof acceleration are contaminated with white Gaussian noise. The variance of the noise $\sigma_{\mathrm{noise}}^2$, the stiffness parameters k_1, k_2, and the damping ratio ξ of the two modes (we assume that the two modes have the same damping ratio) are assumed uncertain, i.e., $\Theta = [\sigma_{\mathrm{noise}}^2, k_1, k_2, \xi]$. The aim of the analysis is to update the belief about the uncertain parameters conditioned on the measurements. The prior distribution is: k_1, k_2 are uniformly distributed on the interval $[0, 3000]$, ξ is uniform on $[0.01, 0.05]$, and $\sigma_{\mathrm{noise}}^2$ is equally likely between $[0, 1]$. The true parameter values used to generate the measured roof acceleration are: $k_1 = k_2 = 1000$, $\xi = 0.03$, $\sigma_{\mathrm{noise}}^2 = 0.2$. The masses are assumed known: $m_1 = m_2 = 1$. For this example problem, we focus on the posterior statistics of k_1, i.e. $g(\Theta) = k_1$.

A reference solution for the problem at hand was obtained by means of rejection sampling [22, 11, 5, 23]. In order to obtain a fairly accurate reference solution, we repeat the procedure of generating 1000 (truly independent) posterior samples more than $7 \cdot 10^4$ times (after that the statistical uncertainty in the estimates is negligible for our purpose). The so-obtained reference values are: the evidence is $c_{E,\mathrm{ref}} \approx 2.817 \cdot 10^{-16}$, the posterior mean of k_1 is $\mu''_{k_1,\mathrm{ref}} \approx 1367$, the posterior standard deviation of k_1 is $\sigma''_{k_1,\mathrm{ref}} \approx 610$.

Example problem 2: bimodal likelihood

As second example problem, we investigate case VII of the second example discussed in [8]: It is a problem with six random variables, where the prior PDF of each random variable is uniform over the interval $[-2, 2]$. The likelihood is a mixture of two multivariate Gaussian distributions centered at $[0.5, \ldots, 0.5]$ and $[-0.5, \ldots, -0.5]$ with a diagonal covariance matrix and standard deviation 0.1; i.e., the likelihood function of this problem is bimodal. Due to the small standard deviation, the two modes are effectively disconnected [8]. $\theta_1, \ldots, \theta_6$ denote the six stochastic variables of the problem. In this example problem we focus on the posterior statistics of $\theta_{\max} = \max(\theta_1, \ldots, \theta_6)$, i.e., $g(\Theta) = \theta_{\max}$.

The probability that a sample from the likelihood is outside of the bounds of the prior distribution is approximately $2.2 \cdot 10^{-50}$. As such an event is negligible, the problem at hand can be solved analytically: $c_E = \frac{1}{4^6} = 2.44 \cdot 10^{-4}$, the posterior mean of $\theta_{\max}$ is $\mu''_{\theta_{\max}} = 0.127$ and the posterior standard deviation of $\theta_{\max}$ is $\sigma''_{\theta_{\max}} = 0.504$. The posterior mean and standard deviation of the parameters θ_i are 0.0 and 0.510.

Example problem 3: sum of Normal random variables

Let Θ be a M-dimensional random vector. The prior distribution of Θ is $f(\Theta) = \prod_{i=1}^{M} \varphi(\theta_i)$, where φ denotes the PDF of the univariate standard Normal distribution. We introduce a function $h(\Theta)$ defined as $h(\Theta) = \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \theta_i$. Note that the prior distribution of $h(\Theta)$ is standard Normal. The likelihood function of the problem is defined through $h(\Theta)$ as $L(\Theta|\mu_\epsilon, \sigma_\epsilon) = \frac{1}{\sigma_\epsilon} \cdot \varphi\left(\frac{h(\Theta) - \mu_\epsilon}{\sigma_\epsilon}\right)$, where $\mu_\epsilon = 4$ and $\sigma_\epsilon = 0.2$. We are particularly interested in the posterior statistics of $h(\Theta)$, thus we set $g(\Theta) = h(\Theta)$.

The problem at hand has an analytical solution: The evidence c_E does not depend on the dimension M of the problem and can be evaluated as $c_E = \frac{1}{\sqrt{1+(\sigma_\epsilon)^2}} \cdot \varphi\left(\frac{\mu_\epsilon}{\sqrt{1+(\sigma_\epsilon)^2}}\right) = 1.785 \cdot 10^{-4}$. The posterior mean and standard deviation of $h(\Theta)$ are $\mu''_h =$

Table 1: Performance of different TMCMC variants (*study 1* & *study 3*) for the bias in the evidence bias_{c_E}, κ_{c_E} and the effective number of independent samples N_{eff}; averaged over 10^4 updating runs with 10^3 posterior samples in each run.

quantities investigated	original TMCMC	TMCMC(β_{opt})	iTMCMC
Example 1			
bias_{c_E}	0.16	0.17	$2 \cdot 10^{-4}$
κ_{c_E}	0.33	0.32	0.34
N_{eff}	11.7	13.7	9.8
bias_{a_g}	$7 \cdot 10^{-3}$	$6 \cdot 10^{-3}$	$8 \cdot 10^{-3}$
bias_{s_g}	0.02	0.01	0.02
Example 2			
bias_{c_E}	0.14	0.42	0.34
κ_{c_E}	2.2	0.89	4.2
N_{eff}	1.8	3.3	2.9
bias_{a_g}	0.03	0.04	0.03
bias_{s_g}	0.01	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$
Example 3 ($M = 6$)			
bias_{c_E}	0.25	0.45	0.11
κ_{c_E}	1.9	0.82	0.59
N_{eff}	1.4	58	70
bias_{a_g}	0.05	$6 \cdot 10^{-3}$	$3 \cdot 10^{-3}$
bias_{s_g}	0.17	0.02	$6 \cdot 10^{-3}$
Example 3 ($M = 100$)			
bias_{c_E}	0.57	0.62	0.63
κ_{c_E}	2.8	2.6	2.6
N_{eff}	0.6	0.9	1.0
bias_{a_g}	0.12	0.09	0.09
bias_{s_g}	0.39	0.30	0.27

$\frac{\mu_\epsilon}{(\sigma_\epsilon)^2+1} = 3.846$ and $\sigma''_h = \sqrt{1/\left(1 + \frac{1}{(\sigma_\epsilon)^2}\right)} = 0.196$, respectively. Note that both the prior and posterior of $h(\boldsymbol{\theta})$ do not depend on the dimension M.

Study 1: Optimal spread of the proposal

In the original TMCMC method, the spread of the proposal distribution is fixed to $\beta = 0.2$. Depending on the problem at hand, this spread can be far from optimal. For this reason, we define an optimal spread β_{opt} of the proposal distribution as the β for which the effective number of independent samples N_{eff} becomes maximal. Thereby, the spread is kept constant in all TMCMC levels. An expensive parameter study is performed to determine the optimal spread β_{opt} associated with each of the example problems. The optimal spread is approximately 0.3 in *Example problem 1*, 0.4 in *Example problem 2*, and $1.3 - 0.18 \cdot \ln(M)$ in *Example problem 3* for $M \in \{6, \ldots, 100\}$. Note that β_{opt} is with respect to a proposal distribution formulated in the underlying standard Normal space (see *proposed modification (3)*), whereas the proposal distribution of the original TMCMC method is expressed in the space spanned by the parameter vector $\boldsymbol{\theta}$, hence they are not directly comparable (except in case of *Example problem 3*).

The performances of the original TMCMC method and the TMCMC variant with the optimal spread β_{opt} are listed in the first two columns of Table 1. For each example problem, the quantities bias_{c_E}, κ_{c_E}, N_{eff}, bias_{a_g} and bias_{s_g} are listed for 10^3 generated posterior samples and 10^3 samples per level. Remember that the optimal spread β_{opt} was selected such that N_{eff} is maximized. Consequently, TMCMC(β_{opt}) clearly outperforms the original TMCMC method with respect to N_{eff}. TMCMC(β_{opt}) performs also better than the original TMCMC method in terms of κ_{c_E}. For the average bias of the estimated evidence bias_{c_E}, the original TMCMC(β_{opt}) method gives a smaller bias. The bias in the estimated posterior mean and standard deviation is very small for both methods and the first two example problems. For the third example problem and different M, the bias in the estimated posterior mean and standard deviation is considerable in both methods. The performance in terms of this example will be investigated in more detail in *study 4*.

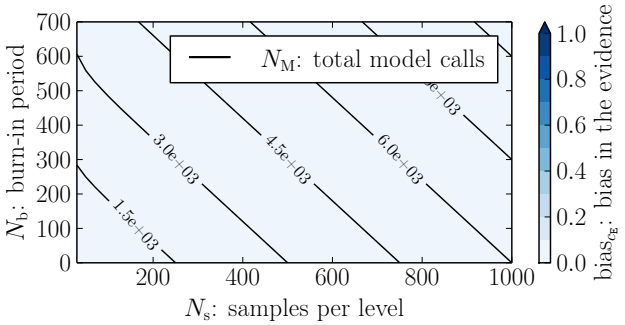
Note that for real problems an optimal spread β_{opt} cannot be determined, since the involved computational burden is much larger than potential advantages.

Study 2: comparison of TMCMC and wTMCMC

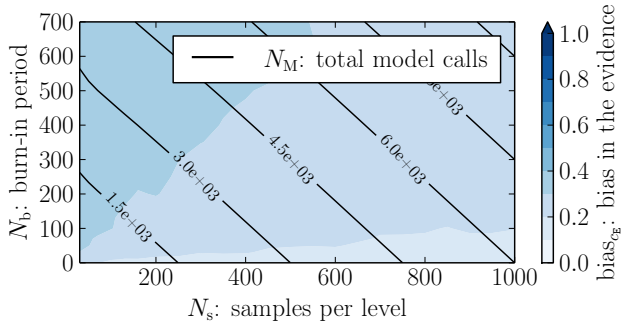
In this study, the TMCMC variant that modifies the sample weights (wTMCMC) is compared to the variant that does not change the weights during the MCMC sampling. We fix the spread of the proposal distribution to β_{opt}, defined in the previous study. Note that β_{opt} is determined for TMCMC (i) without a burn-in phase, (ii) for fixed sample weights, and (iii) with a proposal distribution formulated in the underlying standard Normal space.

The comparison of wTMCMC and TMCMC(β_{opt}) is done for different N_s (the number of posterior samples generated) and N_b (the length of the burn-in period). The number of samples in the intermediate TMCMC-levels is set equal to N_s. In this and all following studies, the results are averaged over 10^4 runs of the updating problem. This number is chosen because it is still computationally feasible and reduces the noise in the presented plots to a tolerable level. In the plots presented as part of this study, the total number N_M of model calls needed to solve the updating problem for given N_s and N_b is indicated. We discuss the performance of both methods with respect to a fixed number of model calls N_M: first we look at the quality of the evidence, and thereafter we assess the quality of the posterior samples.

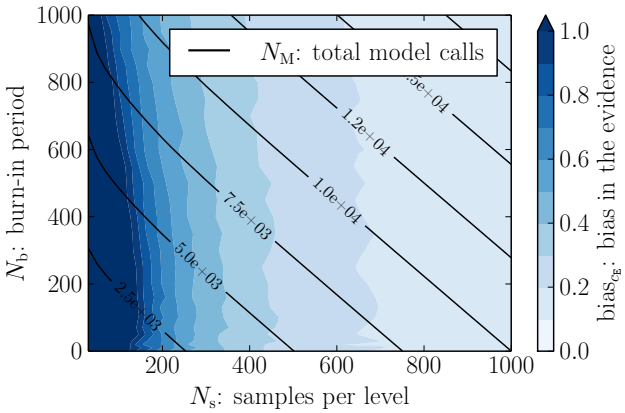
Fig. 2 illustrates the average bias in the estimated evidence, defined according to (Eq. 15). The proposed changes (wTMCMC) reduce the average bias in the estimated evidence considerably for all investigated example problems. For wTMCMC, the bias is mainly controlled by the generated number of posterior samples N_s, and only slightly by the length of the burn-in period N_b. For both wTMCMC and TMCMC(β_{opt}), the smallest bias for a given total number of model calls N_M is obtained using no burn-in, i.e., setting $N_b = 0$. In *Example problem 2*, TMCMC(β_{opt}) method gives zero bias for $N_s \approx 100$ and $N_b = 0$. However, this is only conditional, because the evidence is estimated too large for $N_s < 100$ and too small for $N_s > 100$. Looking at the results of the third example problem for $M = 6$ and $M = 100$ indicates that the bias increases with the dimension of the problem.



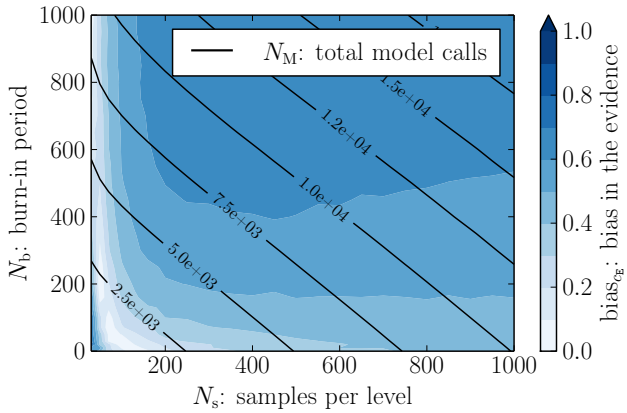
(a) Ex.1: wTMCMC



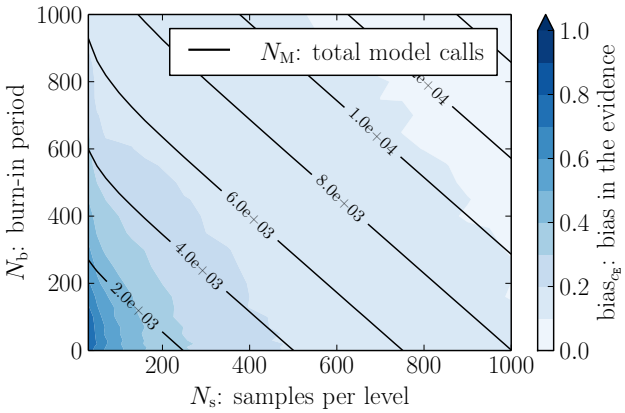
(b) Ex.1: TMCMC(β_{opt})



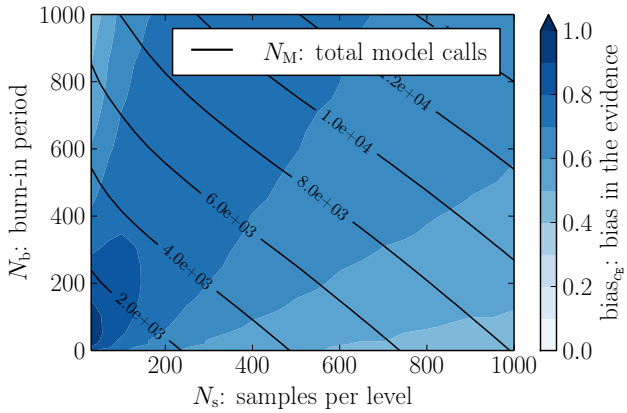
(c) Ex.2: wTMCMC



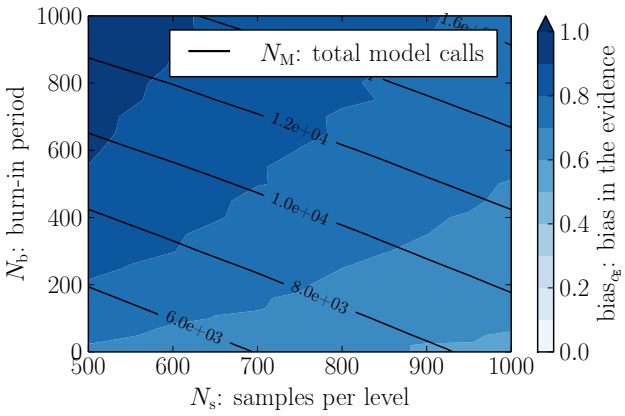
(d) Ex.2: TMCMC(β_{opt})



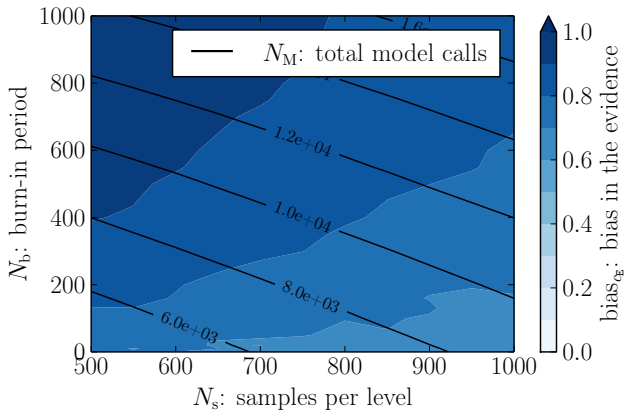
(e) Ex.3 (M = 6): wTMCMC



(f) Ex.3 (M = 6): TMCMC(β_{opt})

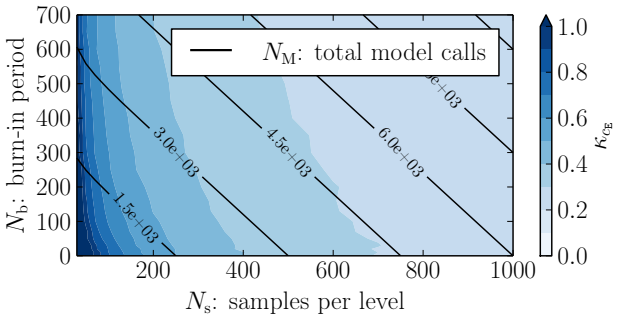


(g) Ex.3 (M = 100): wTMCMC

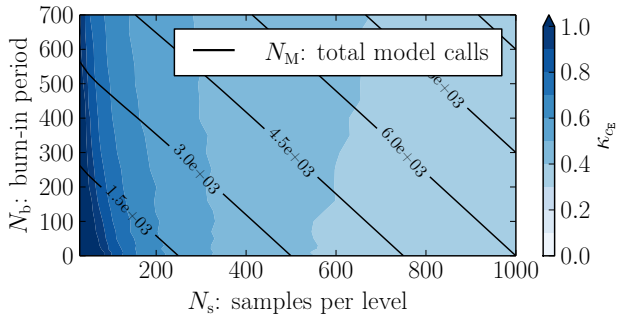


(h) Ex.3 (M = 100): TMCMC(β_{opt})

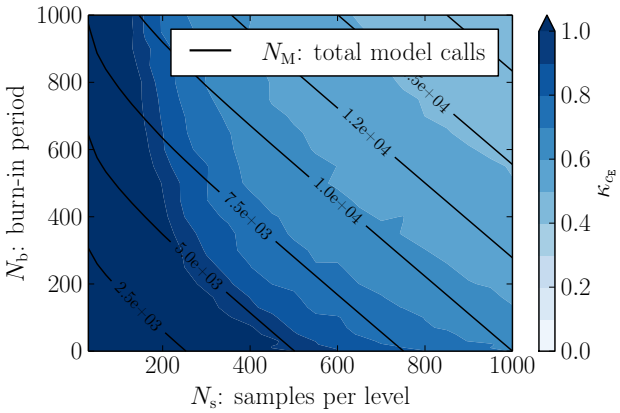
Figure 2: Comparison of wTMCMC and TMCMC(β_{opt}) (*study 2*): bias in the evidence (bias_{c_E})



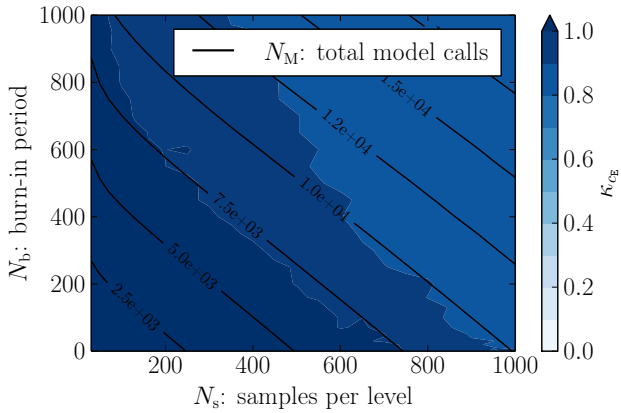
(a) Ex.1: wTMCMC



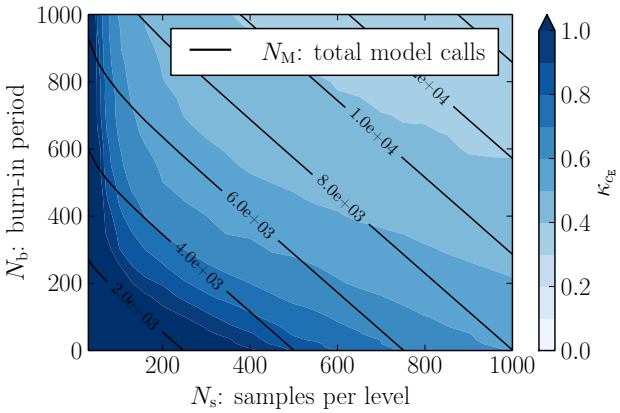
(b) Ex.1: TMCMC(β_{opt})



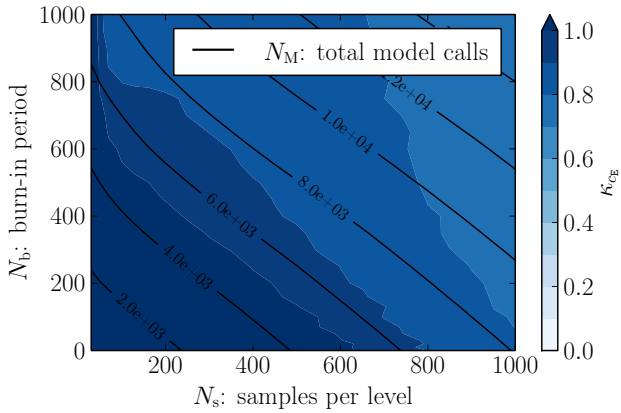
(c) Ex.2: wTMCMC



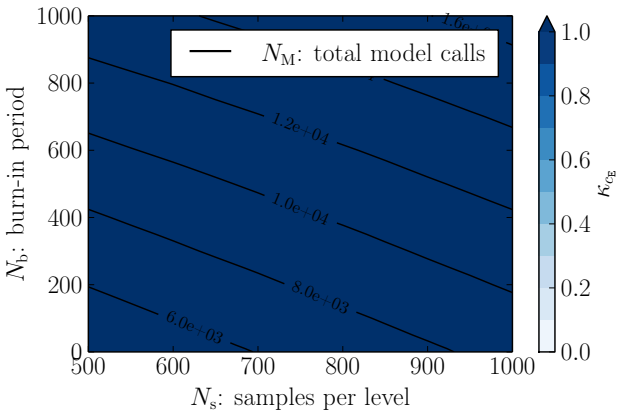
(d) Ex.2: TMCMC(β_{opt})



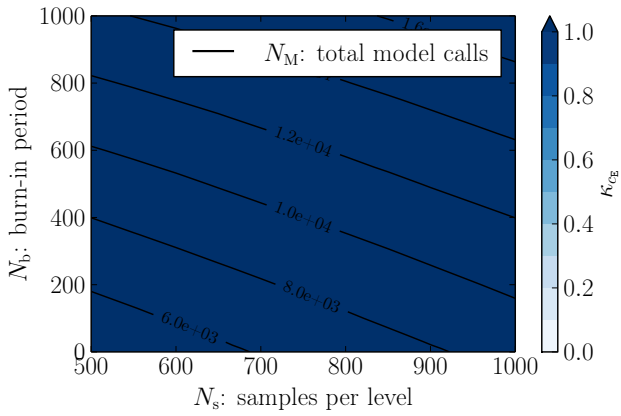
(e) Ex.3 (M = 6): wTMCMC



(f) Ex.3 (M = 6): TMCMC(β_{opt})

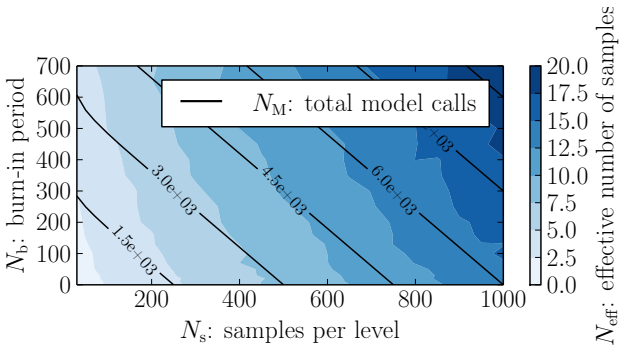


(g) Ex.3 (M = 100): wTMCMC

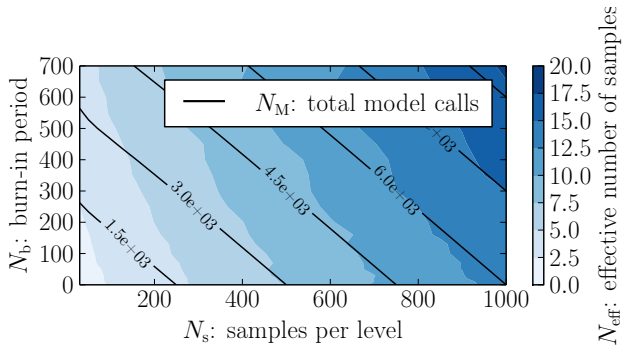


(h) Ex.3 (M = 100): TMCMC(β_{opt})

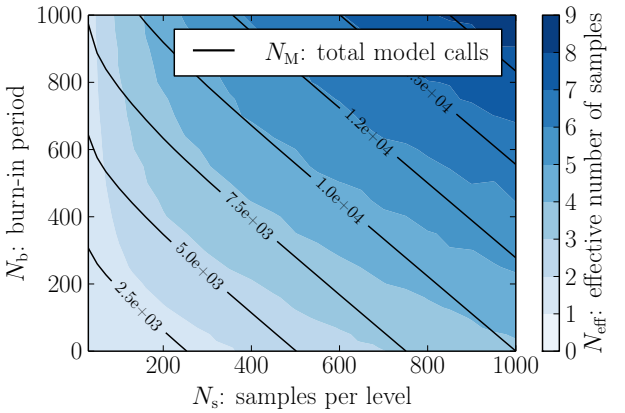
Figure 3: Comparison of wTMCMC and TMCMC(β_{opt}) (*study 2*): κ_{c_E}



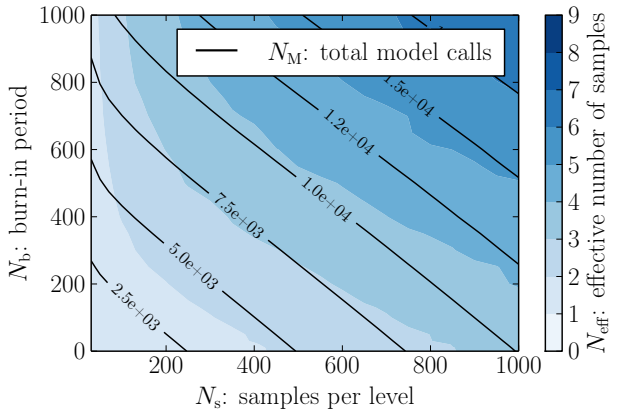
(a) Ex.1: wTMCMC



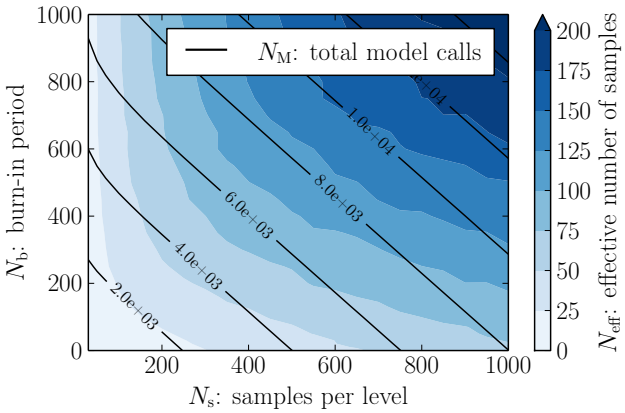
(b) Ex.1: TMCMC(β_{opt})



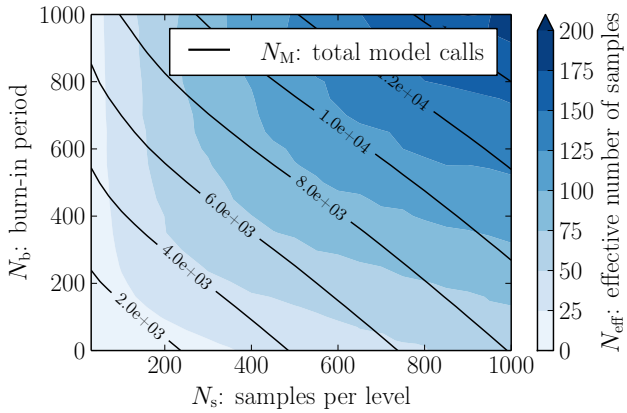
(c) Ex.2: wTMCMC



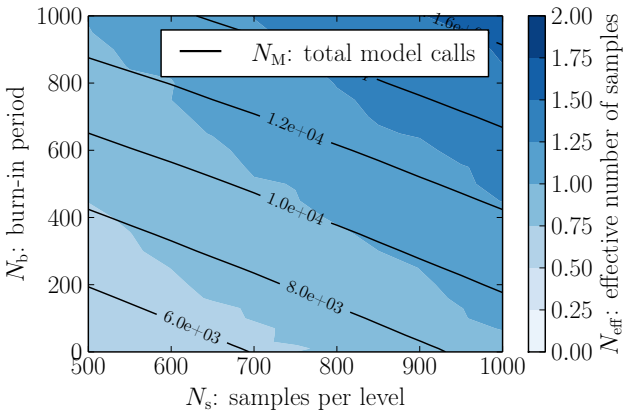
(d) Ex.2: TMCMC(β_{opt})



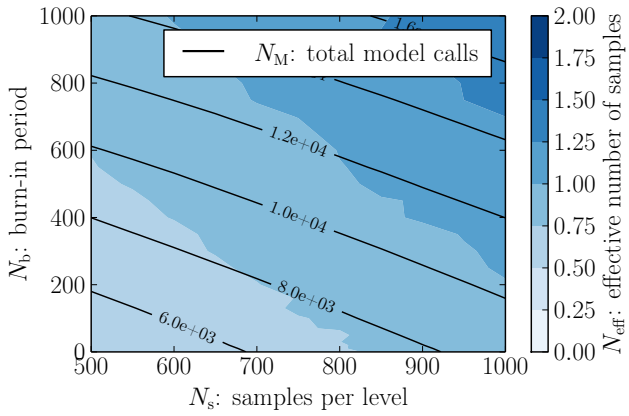
(e) Ex.3 (M = 6): wTMCMC



(f) Ex.3 (M = 6): TMCMC(β_{opt})



(g) Ex.3 (M = 100): wTMCMC



(h) Ex.3 (M = 100): TMCMC(β_{opt})

Figure 4: Comparison of TMCMC(β_{opt}) and wTMCMC (*study 2*): effective number of independent samples N_{eff}.

The average variability of the estimated evidence, see (Eq. 16), is presented in Fig. 3. This measure assesses the overall quality of the estimated evidence, as κ_{c_E} becomes zero only if both the average bias and the variability of the estimated evidence become zero. κ_{c_E} depends on both N_s and N_b for the investigated examples. In this case, a burn-in period can marginally improve κ_{c_E} (for a fixed number of total model calls N_M) – in particular for *Example problem 3* and $M = 6$. However, in practice it is not clear which burn-in period is best for the problem at hand: Therefore, a *zero* burn-in period can be considered a near-optimal solution. Again, wTMCMC performs better than TMCMC(β_{opt}) for the investigated examples. The third example problem demonstrates that an increase of the dimension M leads to larger κ_{c_E}, for both TMCMC(β_{opt}) and wTMCMC.

A study assessing the quality of the estimated posterior mean and standard deviation was performed. The corresponding plots, however, are not shown because they add little to this contribution: With the exception of *Example problem 3* and $M = 100$, we observe the average bias in the estimated mean and standard deviation of the posterior samples to be small for both methods. (as Table 1 already indicates). For reasonable sample sizes (i.e. $N_s > 200$), the bias of both quantities was observed to be smaller than 5%, where wTMCMC performed slightly better than TMCMC(β_{opt}). A notable exception is *Example problem 3* with $M = 100$: In this case a significant bias is observed for both TMCMC(β_{opt}) and wTMCMC. The performance of TMCMC for increasing M in *Example problem 3* is investigated separately in *study 4*.

Fig. 4 depicts the average effective number of independent posterior samples N_{eff}. For the same N_s and N_b, wTMCMC produces a larger effective number of independent samples than TMCMC(β_{opt}). In terms of N_{eff}, a *zero* burn-in period is no longer optimal for all example problems (if the number of total model calls N_M is fixed): For the first example problem and the third example problem with $M = 100$, a *zero* burn-in period is optimal. For the second example problem and the third example problem with $M = 6$, a burn-in period that is approximately as long as N_s produces the largest N_{eff} for a fixed number of total model calls N_M.

To summarize the findings of this study: wTMCMC is observed to outperform TMCMC(β_{opt}), and shows a more consistent convergence behavior in terms of N_s and N_b than TMCMC(β_{opt}). A *zero* burn-in period is found to be optimal for bias$_{c_E}$ and near-optimal for κ_{c_E}, with respect to a fixed number of total model calls N_M. However, a *zero* burn-in period is not optimal in all cases with respect to N_{eff}. As there is no efficient technique to detect an optimal or near-optimal burn-in length for real problems, using a *zero* burn-in period is considered justifiable.

Study 3: Performance of iTMCMC

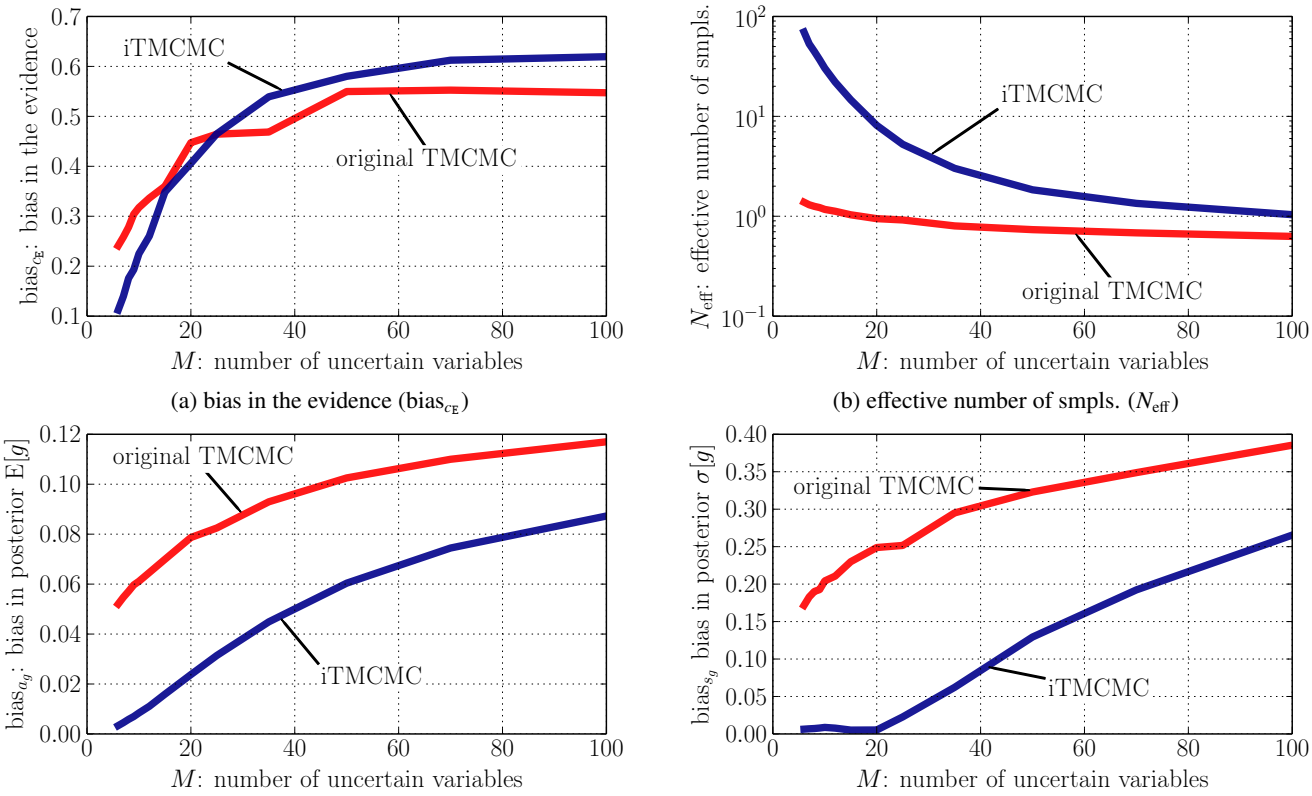
The performance of iTMCMC is compared to the original TMCMC method and to TMCMC(β_{opt}). The results are listed in Table 1. Looking at the bias of the estimated evidence bias$_{c_E}$, iTMCMC outperforms TMCMC(β_{opt}); iTMCMC performs better than the original TMCMC method except for *Example problem 2*. For *Example problem 3* with $M = 100$ all three methods perform similarly bad. Regarding the performance of the investigated methods with respect to κ_{c_E}, a clear tendency cannot be seen. The performance regarding the effective number of independent samples clearly depends on the problem at hand. For *Example problem 1*, iTMCMC is the least efficient among the three investigated TMCMC variants; whereas for *Example problem 3* with $M = 6$, iTMCMC clearly excels. With respect to the bias in the estimated posterior mean and standard deviation, iTMCMC and TMCMC(β_{opt}) perform similarly, and for *Example problem 3* both methods perform better than the original TMCMC method.

In summary: Compared to the original TMCMC method, iTMCMC produces evidence estimates that have on average a smaller bias. By comparing κ_{c_E} with bias$_{c_E}$, we conclude that variability of the evidence estimator increased, while its bias decreased. On average, iTMCMC tends to generate a larger effective number of independent posterior samples and posterior samples with a smaller bias in the mean and standard deviation than the original TMCMC variant. The variant with the optimal spread, TMCMC(β_{opt}), is not an option for practical application, because an optimal spread cannot be found with reasonable computational effort.

Study 4: performance in high dimensions

The performance of iTMCMC and the original TMCMC method for an increasing number of uncertain variables M is investigated by means of *Example problem 3*. The results are depicted in Fig. 5: the average bias in the evidence bias$_{c_E}$, the effective number of independent samples N_{eff}, the bias in the estimate of the posterior mean bias$_{a_g}$ and standard deviation bias$_{s_g}$ is shown in sub-plots (a), (b), (c) and (d), respectively. In a nutshell, the performance of the two methods in terms of the monitored quantities decreases considerably with increasing number of uncertain variables M of the problem. For the investigated *Example problem 3*, the analytical solution of the inference problem does not depend on the dimension M. Therefore, the loss of efficiency with increasing dimension is attributed to the nature of the TMCMC methods. Consequently, care should be taken if a TMCMC method is applied to problems with a large number of uncertain variables. For the example problem at hand, iTMCMC outperforms the original TMCMC method in terms of the bias in the mean and standard deviation and in terms of the effective number of samples. With respect to the bias in the evidence, no clear difference can be observed; for small M iTMCMC is better, and for larger M the original variant is slightly better.

The loss of efficiency in high dimensions has to be attributed mainly to the particular proposal distribution used in TMCMC: it is a multivariate Gaussian distribution with covariance matrix based on the estimated sample covariance. For the multivariate Gaussian proposal distribution, the acceptance/rejection ratio of the Metropolis-Hastings algorithm approaches zero [3] as the dimensionality of the problem increases, rendering the MCMC sampling inefficient. Hybrid proposal distributions can improve the performance of TMCMC in high dimensions, see e.g. [6, 2].



(a) bias in the evidence (bias_{c_E})

(b) effective number of smpls. (N_{eff})

(c) bias in the mean estimate of $h(\mathbf{x})$

(d) bias in the standard deviation of $h(\mathbf{x})$

Figure 5: Performance for an increasing dimension M in *Example problem 3*. (*study 4*)

Summary and Conclusion

The paper proposes three modifications to the algorithm of the original TMCMC method: (1) the TMCMC weights should be adjusted after each MCMC step, (2) a burn-in period may be included for some problems, as the MCMC sampling in TMCMC is only asymptotically *perfect*, (3) the scaling of the proposal distribution can be adjusted adaptively such that a specified target acceptance rate is maintained in the MCMC sampling procedure.

We demonstrated by means of three example problems that: (1) The first proposed modification leads to a better and more consistent convergence behavior (*study 2*). Especially for Bayesian model class selection, the bias of the estimate of the model evidence is reduced by adjusting the TMCMC weights. (2) In most cases, using additional samples is more effective than using a burn-in phase (*study 2*); especially as a optimal burn-in length depends on the problem at hand and is difficult to find. (3) For the third proposed modification, adjusting the spread of the proposal distribution adaptively simplifies the practical application of the method; on average the performance of the method was observed to increase compared to the original TMCMC method (*study 3*). (4) TMCMC works well for problems with few uncertain parameters. However, if the number of uncertain parameters is large, the results can be considerably biased and N_{eff} is very small (*study 4*).

Acknowledgments

With the support of the Technische Universität München – Institute for Advanced Study, funded by the German Excellence Initiative.

References

[1] Andrieu, C., Thoms, J., 2008. A tutorial on adaptive MCMC. Statistics and Computing 18 (4), 343–373.

[2] Angelikopoulos, P., Papadimitriou, C., Koumoutsakos, P., 2015. X-TMCMC: Adaptive kriging for Bayesian inverse modeling. Computer Methods in Applied Mechanics and Engineering 289, 409–428.

[3] Au, S.-K., Beck, J. L., 2001. Estimation of small failure probabilities in high dimensions by subset simulation. Probabilistic Engineering Mechanics 16 (4), 263–277.

[4] Beck, J. L., Yuen, K.-V., 2004. Model selection using response measurements: Bayesian probabilistic approach. Journal of Engineering Mechanics 130 (2), 192–203.

[5] Bolstad, W. M., 2011. Understanding computational Bayesian statistics. Vol. 644. John Wiley & Sons.

[6] Cheung, S. H., 2009. Stochastic analysis, model and reliability updating of complex systems with applications to structural dynamics. Ph.D. thesis, California Institute of Technology.

[7] Cheung, S. H., Beck, J. L., 2009. Bayesian model updating using hybrid Monte Carlo simulation with application to structural dynamic models with many uncertain parameters. Journal of engineering mechanics 135 (4), 243–255.

[8] Ching, J., Chen, Y.-C., 2007. Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. Journal of Engineering Mechanics 133 (7), 816–832.

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

[10] Der Kiureghian, A., Liu, P.-L., 1986. Structural reliability under incomplete probability information. Journal of Engineering Mechanics 112 (1), 85–104.
[11] Gelman, A., Carlin, J. B., Rubin, D. B., Stern, H. S., 2004. Bayesian data analysis.
[12] Gilks, W., Richardson, S., Spiegelhalter, D., 1996. Markov Chain Monte Carlo in practice. CRC press 2.
[13] Hadjidoukas, P., Angelikopoulos, P., Papadimitriou, C., Koumoutsakos, P., 2015. Π4U: A high performance computing framework for Bayesian uncertainty quantification of complex models. Journal of Computational Physics 284, 1–21.
[14] Hohenbichler, M., Rackwitz, R., 1981. Non-normal dependent vectors in structural safety. Journal of the Engineering Mechanics Division 107 (6), 1227–1238.
[15] Jensen, H., Millas, E., Kusanovic, D., Papadimitriou, C., 2014. Model-reduction techniques for Bayesian finite element model updating using dynamic response data. Computer Methods in Applied Mechanics and Engineering 279, 301–324.
[16] MacKay, D. J., 1992. Bayesian interpolation. Neural computation 4 (3), 415–447.
[17] Ortiz, G. A., Alvarez, D. A., Bedoya-Ruíz, D., 2015. Identification of Bouc–Wen type models using the Transitional Markov Chain Monte Carlo method. Computers & Structures 146, 252–269.
[18] Papaioannou, I., Betz, W., Zwirglmaier, K., Straub, D., 2015. MCMC algorithms for subset simulation. Probabilistic Engineering Mechanics 41, 89–103.
[19] Robert, C. P., Casella, G., 2004. Monte Carlo statistical methods, 2nd Edition. Springer.
[20] Roberts, G. O., Gelman, A., Gilks, W. R., et al., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. The annals of applied probability 7 (1), 110–120.
[21] Roberts, G. O., Rosenthal, J. S., et al., 2001. Optimal scaling for various Metropolis-Hastings algorithms. Statistical science 16 (4), 351–367.
[22] Smith, A. F., Gelfand, A. E., 1992. Bayesian statistics without tears: a sampling–resampling perspective. The American Statistician 46 (2), 84–88.
[23] Straub, D., Papaioannou, I., 2015. Bayesian updating with structural reliability methods. Journal of Engineering Mechanics 141 (3).
[24] Wasserman, L., 2000. Bayesian model selection and model averaging. Journal of mathematical psychology 44 (1), 92–107.
[25] Zheng, W., Chen, Y.-T., 2014. Novel probabilistic approach to assessing barge–bridge collision damage based on vibration measurements through transitional Markov chain Monte Carlo sampling. Journal of Civil Structural Health Monitoring 4 (2), 119–131.