Efficiency comparison of MCMC and Transport Map Bayesian posterior estimation for structural health monitoring

Jan Grashorn[a,c,*], Matteo Broggi[a], Ludovic Chamoin[b,c], Michael Beer[a,c,d,e]

[a] *Institute for Risk and Reliability, Leibniz University Hannover, Callinstr. 37, 30167, Hannover, Germany*
[b] *Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS, Laboratoire de Mécanique Paris-Saclay (LMPS), 4 avenue des sciences, 91190, Gif-sur-Yvette, France*
[c] *International Research Training Group 2657, Leibniz University Hannover, Appelstr. 11a, 30167, Hannover, Germany*
[d] *Institute for Risk and Uncertainty, University of Liverpool, Liverpool L69 7ZL 610101, Liverpool, England, United Kingdom*
[e] *International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, Siping road 1239 200092, Shanghai, China*

Abstract

In this paper, an alternative to solving Bayesian inverse problems for structural health monitoring based on a variational formulation with so-called transport maps is examined. The Bayesian inverse formulation is a widely used tool in structural health monitoring applications. While Markov Chain Monte Carlo (MCMC) methods are often implemented in these settings, they come with the problem of using many model evaluations, which in turn can become quite costly. We focus here on recent developments in the field of transport theory, where the problem is formulated as finding a deterministic, invertible mapping between some easy to evaluate reference density and the posterior. The resulting variational formulation can be solved with integration and optimization methods. We develop a general formulation for the application of transport maps to vibration-based structural health monitoring. Further, we study influences of different integration approaches on the efficiency and accuracy of the transport map approach and compare it to the Transitional MCMC algorithm, a widely used method for structural identification. Both methods are applied to a lower-dimensional dynamic model with uni- and multi-modal properties, as well as to a higher-dimensional neural network surrogate system of an airplane structure. We find that transport maps have a significant increase in accuracy and efficiency, when used in the right circumstances.

Keywords: transport maps, Bayesian updating, structural health monitoring, Markov chain Monte Carlo, variational inference

1. Introduction

In most engineering disciplines, data is recorded daily and needs to be integrated into models for parameter estimation, model updating, more robust forecasts of systems, or detection of non-safe system states. Since most sensors are susceptible to noise, stochastic model updating techniques have become a standard to deal with the arising uncertainties. The resulting problem

*Corresponding author, grashorn@irz.uni-hannover.de

is often formulated as a multi-dimensional stochastic problem, where the quantity of interest is the probability distribution of the uncertain parameters. To obtain this distribution, the problem is tackled by using Bayes' theorem, which allows for the formulation of a stochastic distance between model outputs and observed data and the incorporation of a priori available information about the parameters. The resulting posterior distribution often has a complex, unknown topology and a small support in the parameter space, which leads to an ill-defined problem. In the past, many different approaches have been developed to overcome these difficulties, i.e. Laplace estimates [1, 2, 3], variational Bayesian inference [4], Hamiltonian Markov Chain methods [5] and Markov Chain Monte Carlo methods [2, 6]. Specifically in structural problems, the latter have become a standard and since the proposal of the original Metropolis-Hastings (MH) algorithm [7, 8], more efficient alternatives have been developed, i.e. adaptive MH [9], DRAM [10] or Bayesian updating using structural reliability methods (BUS) [11, 12]. See also [13] for a more in-depth summary of sampling methods for Bayesian inverse problems.

The method used in this paper is the Transitional MCMC (TMCMC) method proposed in [14], which introduces a series of intermediate posteriors that are more tractable than the original. By adaptive sampling strategies, the influence of the likelihood can slowly be increased, leading to a much more robust convergence of the Markov chains. TMCMC was also enhanced with adaptive kriging strategies in [15]. Some more notes on the method are given in [16]. MCMC methods have become an important component of parameter inference in structural problems like health monitoring and damage detection, where often measured vibrations, mode shapes, and natural frequencies are used to update parameters in a numerical model. To this extent, [17] developed and used an adaptive MCMC scheme on modal data of a four-story benchmark structure. In [18] MCMC methods were applied to find distribution parameters for long-term data of natural frequencies of an existing bridge. [19] used modal data from a three-story building together with MCMC methods to identify Young's moduli of different sub-structures and their uncertainties. TMCMC specifically was applied to soil characterization uncertainties in [20] and [21], FE model updating with vibration data in [22] and damage detection with ultrasonic wave scattering in [23].

Recent developments in transport theory have opened up a new strategy for variational inference [24, 25], where the goal is shifted from obtaining samples from the posterior to finding a deterministic, invertible mapping from a chosen, easy to calculate, reference density to the posterior density [26]. This transport map (TM) approach has the advantage that once the map is calculated, it allows for the straightforward calculation of integrals on the posterior density and for direct sampling. The mapping is fully deterministic and obtained from an optimization problem which was shown to be unique under certain conditions [26]. Similar developments were undertaken in the field of machine learning with normalizing flows [27, 28, 4, 29], which also construct a mapping between two probability functions. In normalizing flows, the map is considered to be a generative model similar to neural networks [30], while the TM approach mostly uses polynomials [31]. However, there have also been advancements in using tensor trains to decompose the posterior density, thus increasing the efficiency; see [32] and [33] for a discussion and comparison with MCMC methods. We will focus in this paper on the polynomial formulation introduced in [34]. The maps are constructed by using a greedy algorithm which enhances the map order until the desired approximation quality is reached. We aim to compare the TM approach with a proven MCMC method in terms of accuracy and efficiency, and to work out if it is a viable alternative to be used for structural problems. To this extent, we introduce the TM approach with a general formulation for the likelihood under the assumption

of Gaussian noise with an unknown scaling parameter and apply the TM formulation to two problems with varying complexity. The first problem is an analytical mass-spring system, from which the natural frequencies are used to update the spring stiffnesses. The second problem is a neural network surrogate system of a high-dimensional, non-linear airplane test structure (DLRAirmod). Our implementation can be found in [35].
The rest of the paper is structured as follows: In section 2 we will shortly introduce Bayesian updating with MCMC methods, where section 2.1 introduces the general Bayesian formulation, section 2.2 explains the general approach with MCMC methods and section 2.3 shows the transitional MCMC method, which is used in this paper. Section 3 introduces the TM approximation with some general remarks about the implementation. We will give a short summary of the differences in MCMC and TM approaches in section 3.4. In section 4, two numerical examples are shown, where we focus on the efficiency based on several parameters of the approaches and the posterior. Section 4.1 shows different integration schemes of the TM approach used on a lower-dimensional dynamic model, while section 4.2 shows the application to a neural network surrogate model of a finite element airplane structure. Section 5 concludes this paper.

2. Parameter Estimation

2.1. Bayesian framework

Let $\boldsymbol{\theta} \in \mathbb{R}^d$ be a d-dimensional random variable with prior probability $p(\boldsymbol{\theta})$ describing uncertain parameters of a model $\mathcal{M}(\boldsymbol{\theta})$, where $\mathcal{M}(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}^n$. Given measured data $\boldsymbol{D} \in \mathbb{R}^{n\times m}$ the probability of observing $\boldsymbol{\theta}$ in $\mathcal{M}(\boldsymbol{\theta})$ under the condition of $\boldsymbol{D}$ can be calculated using Bayes' theorem [1]

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{D})} \tag{1}$$

where the likelihood $p(\boldsymbol{D}|\boldsymbol{\theta})$ describes the probability of observing the data under the assumption of $\boldsymbol{\theta}$ and is usually modeled as a stochastic distance between $\mathcal{M}(\boldsymbol{\theta})$ and $\boldsymbol{D}$. The most common assumption here is a Gaussian distribution centered on $\mathcal{M}(\boldsymbol{\theta})$. With some zero-mean additive noise parameter $\boldsymbol{\zeta} \sim \mathrm{N}(0, \boldsymbol{\Sigma}_\zeta)$ the model output $\boldsymbol{y} \in \mathbb{R}^n$ becomes

$$\boldsymbol{y} = \mathcal{M}(\boldsymbol{\theta}) + \boldsymbol{\zeta} \tag{2}$$

and the likelihood can be modeled as a Gaussian with mean $\boldsymbol{y}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}$ [1]. The obtained posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{D})$ is an expression for the updated probability for $\boldsymbol{\theta}$ constrained by the observation of $\boldsymbol{D}$. The so-called evidence $p(\boldsymbol{D})$ is constant for any given set of model and data so Eq. (1) is also used in the non-normalized form

$$p(\boldsymbol{\theta}|\boldsymbol{D}) \propto p(\boldsymbol{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \ . \tag{3}$$

While this prohibits the direct evaluation of the posterior density, the formulation in Eq.(3) drastically reduces the complexity for sampling or variational methods, since calculating $p(\boldsymbol{D})$ involves the integration over the full likelihood, i.e.

$$p(\boldsymbol{D}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \ . \tag{4}$$

2.2. MCMC-based inference

One key difficulty in Bayesian model updating (BMU) is the irregular and unknown shape of $p(\boldsymbol{\theta}|\boldsymbol{D})$ and the fact that it can only be evaluated point-wise. Therefore, MCMC methods are usually employed to explore the probability space based on random walks and extract statistical information, e.g. means and variances of the updated parameters [6]. The employed strategies most often construct Markov chains based on some form of rejection sampling. Given some state $\boldsymbol{\theta}_k$ and a proposal state $\hat{\boldsymbol{\theta}}_k$ drawn from a proposal distribution p_k, the proposal state is accepted as $\boldsymbol{\theta}_{k+1}$ with probability

$$p_k = \min\left(1, \frac{p(\hat{\boldsymbol{\theta}}_k)}{p(\boldsymbol{\theta}_k)}\right), \tag{5}$$

otherwise $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$. Different strategies have been proposed to sample $\hat{\boldsymbol{\theta}}_k$, i.e. the original MH implementation [7, 8] chooses a global proposal distribution, whereas the adaptive MH-algorithm [9] centers the proposal distribution around the current sample $\boldsymbol{\theta}_k$. MCMC methods can introduce non-ergodicity in cases of multi-modal or narrowly supported target distributions. It is difficult for the Markov chains to jump from one mode to another, if the proposal distribution is chosen to be too narrow. However, if it is too wide, many proposal states will be rejected, leading to a waste of computation time. The choice of an adequate proposal distribution is therefore important for the efficiency and accuracy of MCMC methods. If the posterior's support is not known a-priori, there is also often the need for burn-in samples, which further increases the computational costs.

2.3. Transitional MCMC

The main idea of the TMCMC scheme [14] used in this paper is to introduce an exponent $\alpha_j \in [0, 1]$ to the likelihood

$$p(\boldsymbol{\theta}|\boldsymbol{D}) \propto p(\boldsymbol{D}|\boldsymbol{\theta})^{\alpha_j}\, p(\boldsymbol{\theta}) \tag{6}$$

and increasing α_j with each level j starting from $\alpha_1 = 0$, which is equal to sampling from the prior density. Note that for $\alpha_j = 1$, Eq. (6) becomes Eq. (3), therefore samples at the last level are distributed according to the posterior. Values of α_j for the intermediate levels are chosen based on the variance of the drawn samples. After drawing samples from the prior density, the Adaptive Metropolis-Hastings algorithm [9] is used to draw samples for the next level until $\alpha_j = 1$ is reached. The main motivation behind TMCMC is to avoid the problem of sampling from difficult target probability density functions (PDFs) but sampling from a series of PDFs that converge to the target PDF and that are easier to sample [14]. Through this process it is also possible to sample from multi-modal, narrow posteriors since the samples are "pushed" slowly towards regions of higher probability. The need for burn-in samples is also removed, since samples at each level j are guaranteed to be drawn according to $p(\boldsymbol{D}|\boldsymbol{\theta})^{\alpha_j}\, p(\boldsymbol{\theta})$.

3. Transport Maps

A transport map M is a deterministic coupling between a reference density ρ and the target density π

$$\int_Y f(y)\pi(y)\mathrm{d}y = \int_X f(M(x))\rho(x)\mathrm{d}x \tag{7}$$

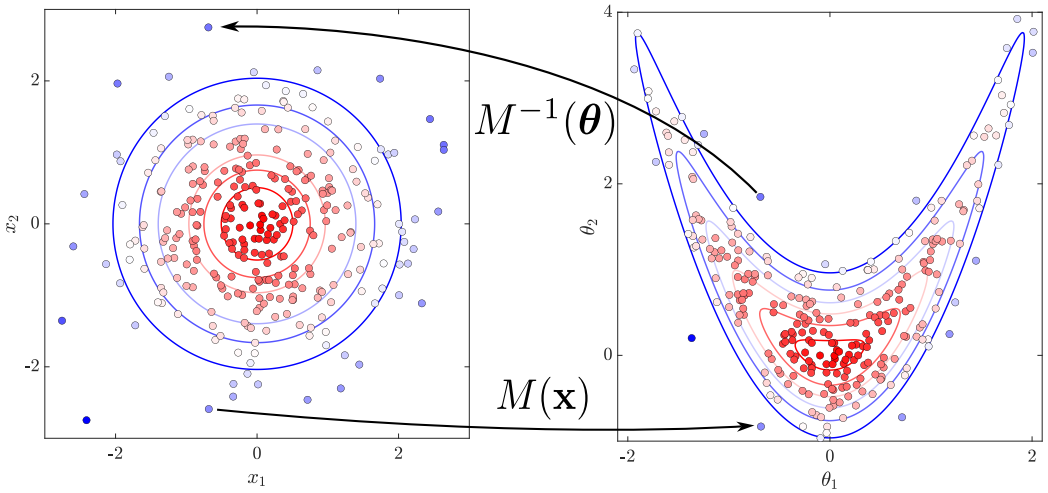


Figure 1: Example of mapping for $\boldsymbol{\theta} \in \mathbb{R}^2$ between standard-normal reference density (left) and target density (right). The map $M(\boldsymbol{x})$ pushes samples $\boldsymbol{x}$ forward to the target density, $M^{-1}(\boldsymbol{\theta})$ pulls them back to the reference density.

where the target density in the case of BMU is the posterior distribution. The reference density can be chosen freely by the analyst. Common choices are standard normal or standard uniform distributions [31]. Any integrals on the target density can thus be calculated on the reference density by use of the map M in conjunction with classical quadrature rules. Moreover, samples from the target density Y can be drawn by drawing samples X from the reference density and then evaluating the map M. This makes it possible to find an analytical formulation for the posterior density in BMU, which is usually difficult or impossible. The task now becomes to find the map M. A map can be any invertible function $M : \mathbb{R}^d \to \mathbb{R}^d$, e.g. polynomials or even neural networks [36]. Note that in the context of this paper maps M are defined to transport samples from ρ to π, its inverse M^{-1} then transports samples from π to ρ. Due to invertibility, the direction of the transport is somewhat arbitrary and usually depends on the application, since with transport maps it is possible to either approximate densities from an evaluable expression (as is the case for BMU) or approximate a density from samples of the target [36]. Furthermore, we define samples in the target space as $\boldsymbol{\theta}$ and samples in the reference space as $\boldsymbol{x}$. An illustration of the transport process is shown in Fig. 1. Using the notation $M_\#$ for the push-forward operation

$$M_\#\rho = \pi(M(\boldsymbol{x})) \cdot |\det \nabla M(\boldsymbol{x})| \tag{8}$$

and $M^\#$ for the pull-back

$$M^\#\pi = \rho(M^{-1}(\boldsymbol{\theta})) \cdot |\det \nabla M^{-1}(\boldsymbol{\theta})| \tag{9}$$

the mismatch of the approximation $\pi \approx M_\#\rho$ can be expressed with the Kullback-Leibler (KL) divergence

$$\mathcal{D}_{\mathrm{KL}}(M_{\#}\rho \,||\, \pi) = \mathcal{D}_{\mathrm{KL}}(\rho \,||\, M^{\#}\pi) \tag{10}$$

$$= \mathbb{E}_{\rho}\Big[\log \frac{\rho}{M^{\#}\pi}\Big] \tag{11}$$

where the invertibility of the map is used in Eq. (10). With $\boldsymbol{a}$ as map parameters, Eq. (11) becomes

$$\mathcal{D}_{\mathrm{KL}}(M_{\#}\rho||\pi) = \int_{\boldsymbol{X}} \Big[\log \rho(\boldsymbol{x}) - \log \pi(M(\boldsymbol{a},\boldsymbol{x})) - \log[|\det \nabla M(\boldsymbol{a},\boldsymbol{x})|]\Big]\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \ . \tag{12}$$

Due to optimality and uniqueness properties, as well as being computationally practical, maps M were proposed to be monotonic, lower-triangular, and constructed from components [34]

$$M^k(\boldsymbol{x}_{1,...,k}, \boldsymbol{a}) = f(x_1, ..., x_{k-1}, 0, \boldsymbol{a}) + \int_0^{x_k} g(\partial_k \ f(x_1, ..., x_{k-1}, \bar{x}, \boldsymbol{a}))d\bar{x} \tag{13}$$

where $f : \mathbb{R}^d \to \mathbb{R}$. $g : \mathbb{R} \to \mathbb{R}_+$ is a rectifier, e.g. the exponential function, ensuring monotonicity. The resulting map has the structure

$$M(\boldsymbol{x}) = \begin{bmatrix} M^1(x_1) \\ ... \\ M^d(x_1, ..., x_d) \end{bmatrix} . \tag{14}$$

Moreover, each M^k is a function $M^k : \mathbb{R}^k \to \mathbb{R}$, the full map $M(\boldsymbol{x})$ is thus a function $M(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^d$. The formulation of Eq. (13) ensures that the component M^k is monotonic at least in the k-th component [34]. The functions f are defined as

$$f(x_1, ..., x_k, \boldsymbol{a}) = \sum_{\boldsymbol{\alpha} \in \boldsymbol{\Lambda}} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(x_1, ..., x_k) \tag{15}$$

with scalar coefficients $a_{\boldsymbol{\alpha}} \in \mathbb{R}$ and a set $\boldsymbol{\Lambda}_{m \times d}$ of m multi-indices $\boldsymbol{\alpha}_{1\times d}$, where each entry $(\boldsymbol{\alpha}_i)_j \in \mathbb{N}_0$ describes the function degree for the i-th multi-index in the j-th input. The basis functions $\Psi_{\boldsymbol{\alpha}}$ are constructed from the tensor product of univariate polynomials according to the multi-indices

$$\Psi_{\boldsymbol{\alpha}}(x_1, ..., x_k) = \prod_{i=1}^{k} \Psi_{\alpha_i}(x_i) \ . \tag{16}$$

While the maps could theoretically be constructed from any function, constructing them in this fashion has significant advantages in terms of uniqueness and invertibility. The triangular structure of Eq. (14) leads to a well-defined optimization problem and a unique map while also ensuring that the maps remain computationally tractable. Invertibility, although not strictly necessary, is needed for the computation of integrals in the target density. For more details and proofs we refer to [34].

A good approximation of the posterior density results in a small KL divergence, so that Eq. (12) can be transformed into a minimization problem to calculate the map parameters $\boldsymbol{a}$. Note that

$\rho(x)$ does not depend on the map parameters, and instead of the full posterior $\pi = p(\boldsymbol{\theta}|\boldsymbol{D})$, the non-normalized form

$$\tilde{\pi} = p(\boldsymbol{D}|\boldsymbol{\theta})\ p(\boldsymbol{\theta}) \tag{17}$$

can be used. Furthermore, since M consists of analytical functions, the involved integrals in Eq. (13) can easily be computed by suitable quadrature rules. Moreover, the integral in Eq. (12) can also be approximated by quadrature rules, i.e. Gauß or Monte Carlo. A short discussion on different methods is given in section 3.1. Setting $w_{q,i}$ and $\boldsymbol{x}_{q,i}$ as weights and integration points for the quadrature rule in Eq. (12), the final minimization problem to obtain the needed map parameters $\boldsymbol{a}$ is then

$$\min_{\boldsymbol{a}} \sum_i w_{q,i} \Big[-\log\Big(\widetilde{\pi}(M(\boldsymbol{a}, \boldsymbol{x}_{q,i}))\Big) - \log(|\det \nabla M(\boldsymbol{a}, \boldsymbol{x}_{q,i})|)\Big] \ . \tag{18}$$

Eq. (18) can also be viewed as maximizing the integral of the mapping.

For computations, the framework MParT [37] was used which handles most of the map computations, specifically the expressions in Eqs. (13) and (14). As a basis for the functions f in Eq. (13) Hermite polynomials up to a degree n are chosen. Higher order functions are usually able to approximate the target density with higher accuracy, however, since more parameters need to be optimized their computation is also more costly. Since the needed map order cannot be known a priori and to reduce the computational burden, it is useful to formulate an adaptive algorithm that successively adds components to the map as long as a user-defined tolerance for the approximation quality is not reached [34]. This tolerance has some implications to the resulting approximation and the efficiency of the procedure. A higher tolerance leads to worse performance while a lower tolerance increases the computational effort. Throughout this paper we chose a value of 10^{-4} as we found this a good compromise between cost and performance. The component order is increased step-wise, i.e. to increase the order from a map M_{Λ_m} with order m to order $m+1$ all components $\Lambda_{m+1} \backslash \Lambda_m$ are included. Note that more sophisticated techniques were proposed, e.g. using underlying sparsity patterns. For an in-depth discussion the reader is referred to [31, 34]. A greedy algorithm (see Algorithm 1) can be constructed such that only the new map components need to be optimized.

A key aspect of transport maps is solving the optimization procedure (18) for which a variety of algorithms exist. To increase efficiency, algorithms that utilize gradient information can be employed, however, this ultimately means that the posterior density and by that also the model has to be differentiable with respect to the parameters. This is rarely the case in engineering applications where numerical models (i.e. FEM) are used. One way around this issue is to use reduced order modeling or surrogate models based on differentiable functions, e.g. polynomial chaos expansion, proper generalized decomposition, or neural networks [38, 39].

3.1. Quadrature rule in the optimization

The choice of the quadrature rule in Eq. (18) plays a significant role in the efficiency and accuracy of the transport map approach. In the optimization, each integration point is mapped to the target space, where the target function is then evaluated. This means that the number of integration points should be kept low to avoid too many model evaluations. However, the number of points needs also to be chosen such that the accuracy is not suffering. Because the reference density is easily computable, some quadrature schemes apply here [32]. We investigated

Algorithm 1 Adaptive map generation

# Construct first map component
order = 1
Λ_1 = GetMultiIndices (order)
map = CreateMap (Λ_1)
Minimize ($\mathcal{D}_{\mathrm{KL}}$(map, $\tilde{\pi}$)) (See Eq. (18))

# Increase map order while KL-divergence is too high
while $\mathcal{D}_{\mathrm{KL}}$(map, $\tilde{\pi}$) > TOL **do**
    order += 1
    Λ_{order} = GetMultiIndices (order)
    $\Lambda_{\text{order}} = \Lambda_{\text{order}} \backslash \Lambda_{\text{order-1}}$ # retain only the new indices
    map = Compose (map, CreateMap (Λ_{order})) # create new map from old map and new components
    Minimize ($\mathcal{D}_{\mathrm{KL}}$(map, $\tilde{\pi}$))
end while

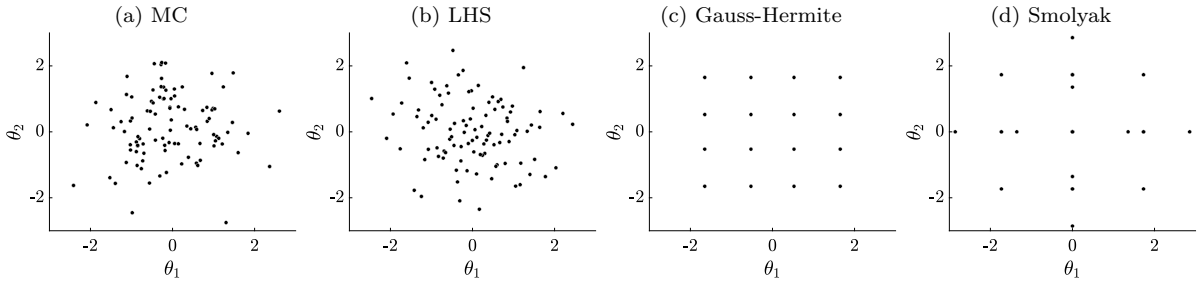


Figure 2: Example of integration points for $\boldsymbol{\theta} \in \mathbb{R}^2$, a): MC (100 points), b): LHS (100 points), c): Gauss-Hermite (16 points), d):Smolyak (25 points)

two random sampling approaches, namely Monte Carlo (MC) and Latin hypercube sampling (LHS), as well as Gauss-Hermite integration and sparse grids with Smolyak's algorithm [40, 41]. Examples for each of these algorithms can be seen in Figure 2. It is evident that in Gauss-Hermite and Smolyak integration a lower number of points can cover the standard normal space far more efficiently than any of the Monte Carlo schemes.

3.2. Laplace approximation

As a first approximation a map to a Gaussian with mean $\boldsymbol{\theta}_0$ and covariance Σ_0 can be calculated by using the Laplace approximation. The role of the Laplace approximation is essentially to shift the target density function to the origin and to normalize the covariance. This results in finding the best local Gaussian approximation. It is also similar to finding the maximum a-posteriori (MAP) estimate. $\boldsymbol{\theta}_0$ and Σ_0 are found by solving the optimization problem

$$\boldsymbol{\theta}_0 = \arg\min_{\theta} \quad -\log\tilde{\pi}(\boldsymbol{\theta}) \tag{19}$$

which corresponds to finding the mode of the posterior. Σ_0 corresponds to finding the Hessian $\mathcal{H}$ at $\boldsymbol{\theta}_0$, the final Laplace map thus becomes [38]

$$L(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 - \mathcal{H}^{-\frac{1}{2}}\boldsymbol{\theta} \ . \tag{20}$$

The Laplace approximation can be used to regularize the problem when combined with maps M to give the final form of $\mathbb{M}$

$$\mathbb{M}(\boldsymbol{x}) = L \circ M(\boldsymbol{x}) \ . \tag{21}$$

Figure 3 shows the transport process with Laplace approximation and with an intermediate transport map on a two-dimensional example. Note that the intermediate density has zero mean and the resulting density approximations have roughly the same mean and covariance structure. The full algorithm to create the map $\mathbb{M}(\boldsymbol{x})$ is shown in Algorithm 2.

3.3. Accuracy

As a measure for the accuracy of the TM-approximation, the variance diagnostic introduced in [25] can be used

$$\varepsilon_\sigma = \frac{1}{2}\mathbf{Var}\Big[\log\tilde{\pi}(\mathbb{M}(\boldsymbol{x})) + \log|\det\nabla\mathbb{M}(\boldsymbol{x})| - \rho(\boldsymbol{x})\Big] \ . \tag{22}$$

The evaluation of Eq. (22) does not come with additional computational costs since it is directly evaluated in the optimization procedure (18). With this measure, the approximation quality can be assessed, and if deemed too low the map order can be increased adaptively.

3.4. Key aspects of MCMC and TM

The main difference between the MCMC and TM approaches is the fact that MCMC methods obtain a collection of samples from the posterior, while TM directly approximates the underlying density in a deterministic fashion. This coupling allows for direct sampling of the posterior, as well as the calculation of integrals. MCMC methods obtain samples by a "local" approach, i.e. they draw samples according to the probabilites of the current and a proposal sample and do not take into account the global structure of the posterior. Through the use of the Kullback-Leibler

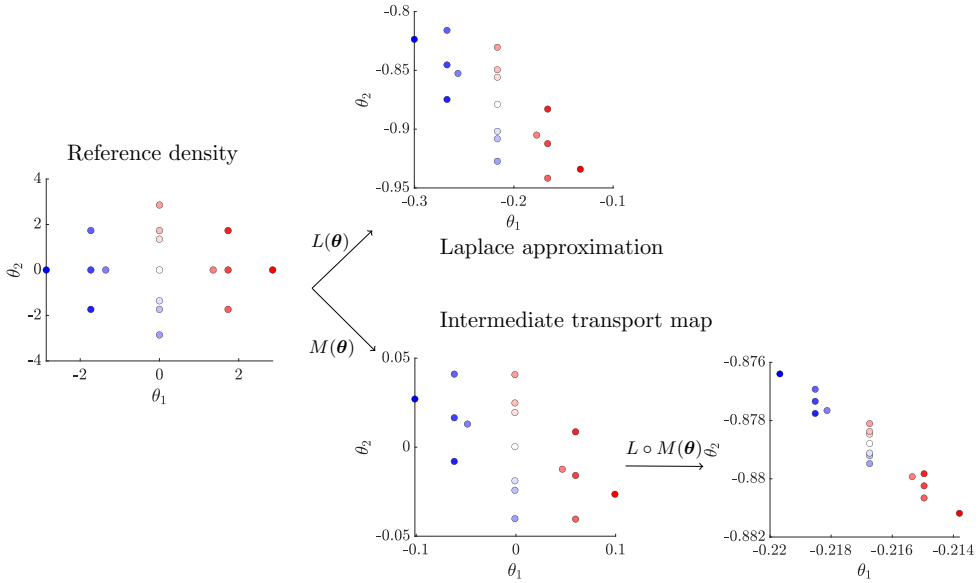


Figure 3: Example of transport of Smolyak integration points from standard normal reference distribution with pure Laplace approximation (top) and intermediate transport map (bottom).

Algorithm 2 Adaptive map generation with Laplace map

Construct Laplace map
$\boldsymbol{\theta}_0$ = Min (- log ($\tilde{\pi}(\boldsymbol{\theta})$))
$\mathcal{H}$ = Hessian ($\tilde{\pi}(\boldsymbol{\theta}_0)$)
L = LaplaceMap (θ_0, $\mathcal{H}$)

# Construct first map component
order = 1
Λ_1 = GetMultiIndices (order)
map = CreateMap (Λ_1)
map = Compose (L, map)
Minimize($\mathcal{D}_{\mathrm{KL}}$(map, $\tilde{\pi}$)) (See Eq. (18))

# Increase map order while KL-divergence is too high
while $\mathcal{D}_{\mathrm{KL}}$(map, $\tilde{\pi}$) > TOL **do**
    order += 1
    Λ_{order} = GetMultiIndices (order)
    $\Lambda_{\text{order}} = \Lambda_{\text{order}} \backslash \Lambda_{\text{order-1}}$ # retain only the new indices
    map = Compose (map, CreateMap (Λ_{order})) # create new map from old map and new components
    Minimize ($\mathcal{D}_{\mathrm{KL}}$(map, $\tilde{\pi}$))
end while

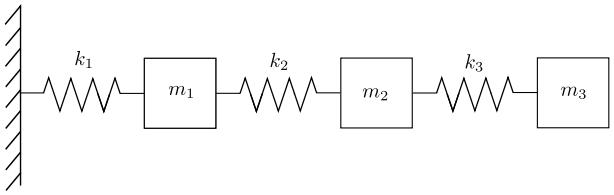


Figure 4: 3-DOF model with $m_1 = m$, $m_2 = 2m$ and $m_3 = 3m$.

divergence as optimization target, transport maps theoretically have global information about the posterior. However, as we will show in the next section, there are some issues associated with this, especially in multi-modal posteriors. While these usually pose a challenge, especially due to the limited knowledge of the target space, some techniques are available to deal with the arising issues. TMCMC makes use of so-called annealing [42], where instead of directly approximating the target density, it is estimated through a series of intermediate densities. These allow for a slower convergence to the density of interest, such that multiple modes are more likely to be captured. This concept was recently shown to work with transport maps [32, 43]. Instead of calculating a single direct map, a series of maps is chained together. Each map thus transports samples between some intermediate densities that converge to the target. The interested reader is also referred to [44, 45].
Generally, implementation of MCMC is more straight-forward, since the model can be used in a black-box fashion and the updating is performed directly on the model parameters. Transport maps require more involved calculations, as an additional set of parameters is needed. For the TM approach, gradients are often used to increase the optimization efficiency. Here it is possible to use recent developments for surrogate or reduced order models based on analytical functions, such as proper orthogonal decomposition (POD), proper generalized decomposition (PGD) or neural networks (see e.g. [38] for an application with PGD). Note that there are also MCMC algorithms that rely on gradients, such as the No-U-Turn sampler (NUTS, [46]). Moreover, there exist a variety of methods for gradient-free optimization of the maps to be used in cases where model-gradients are not available. Finally, given a suitable framework, transport maps have less dependency on user-chosen parameters, such as the proposal distribution, number of burn-in samples or the number of Markov chains.

4. Numerical examples

We test the accuracy and efficiency of both TMCMC and TM methods on two examples, one being a 3-DOF dynamical system with four springs and the other one an FE-model of an airplane structure. We also conduct a study on the influence of different integration schemes for solving Eq. (18).

4.1. Dynamical model

The dynamical model consists of three masses ($m_1 = m$, $m_2 = 2m$, $m_3 = 3m$) and three springs ($k_{1,\dots,3}$). Consider $\tilde{\boldsymbol{\theta}}$ as the random parameters in physical space of the model. The goal is to update the probability densities for the springs in two cases:

Case I: $\quad \tilde{\boldsymbol{\theta}} = \begin{bmatrix} k_1 & \sigma \end{bmatrix}^\top$

Case II: $\tilde{\boldsymbol{\theta}} = \begin{bmatrix} k_1 & k_2 & k_3 & \sigma \end{bmatrix}^\top$

based on calculations of the three natural frequencies $\boldsymbol{\omega}$. Using $\boldsymbol{K}$ as the stiffness matrix and $\boldsymbol{M}$ as the mass matrix

$$\boldsymbol{K} = \begin{bmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} m & 0 & 0 \\ 0 & 2m & 0 \\ 0 & 0 & 3m \end{bmatrix} \tag{23}$$

$\boldsymbol{\omega}$ is calculated through solving the eigenvalue problem

$$\det(\boldsymbol{K}(\tilde{\boldsymbol{\theta}}) \cdot \boldsymbol{M}^{-1} - \boldsymbol{\omega}\boldsymbol{I}) = 0 \ , \tag{24}$$

i.e. $\mathcal{M}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{\omega}$. We write $\boldsymbol{K}(\tilde{\boldsymbol{\theta}})$ since the stiffness is dependent on the random parameter $\tilde{\boldsymbol{\theta}}$. Measurements were obtained from 50 Monte Carlo simulations by assuming randomness in all springs with $\mu_k = [1500, 750, 1200]$ and a variance of 1% of the mean to simulate noise in the data. The data is additionally corrupted by an additive noise with a standard deviation of again 1% of the measured values. In case I the values for k_2 and k_3 were set to their respective mean values in the updating and not considered to be random. As a prior we use a uniform distribution $\tilde{\theta}_i \sim \mathrm{U}(100, 3500)$ for the uncertain stiffness parameters. To make calculations easier all of the calculations apart from the model are performed in standard normal space, such that the random variables need to be transformed to physical space

$$\tilde{\theta}_i = 100 + \Phi(\theta_i) \cdot (3500 - 100) \ . \tag{25}$$

$\Phi(\cdot)$ denotes the standard normal cumulative density function (CDF), θ_i is the i-th component of $\boldsymbol{\theta}$ in standard normal space. The prior in standard normal space is then the standard normal density. Assuming a Gaussian noise model with covariance matrix $\boldsymbol{\Sigma} = \mathrm{cov}[\boldsymbol{D}]$ and a scaling factor σ the log-posterior and its derivative are calculated by

$$\log p(\boldsymbol{\theta}|\boldsymbol{D}) = -m \log\left(\sqrt{(2\pi)^{|\boldsymbol{D}|} \cdot \sigma \det \boldsymbol{\Sigma}}\right) - \sum_{i=1}^{m} (\boldsymbol{D}_i - \mathcal{M}(\boldsymbol{\theta}))^\top \cdot \frac{1}{\sigma} \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{D}_i - \mathcal{M}(\boldsymbol{\theta})) - 0.5 \sum_i \theta_i^2 \tag{26}$$

$$\frac{\partial \log p(\boldsymbol{\theta}|\boldsymbol{D})}{\partial \boldsymbol{\theta}} = \frac{\partial p(\boldsymbol{D}|\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}}} \cdot \frac{\partial \tilde{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} - \boldsymbol{\theta} \tag{27}$$

$$= \frac{\partial p(\boldsymbol{D}|\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}}} \cdot (3500 - 100)\, \phi(\boldsymbol{\theta}) - \boldsymbol{\theta} \tag{28}$$

where m is the dimensionality of the data, $|\boldsymbol{D}|$ the cardinality of the data (in this case $|\boldsymbol{D}| = 3$), $\phi(\cdot)$ the standard normal density function and the derivative of the likelihood function $\partial p(\boldsymbol{D}|\boldsymbol{\theta})/\partial \tilde{\boldsymbol{\theta}}$ is calculated with

$$\frac{\partial p(\boldsymbol{D}|\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}}} = -\frac{1}{\sigma^2} \sum_i^m (\boldsymbol{D}_i - \mathcal{M}(\tilde{\boldsymbol{\theta}}))^\top \cdot \boldsymbol{\Sigma}^{-1} \cdot \frac{\partial \mathcal{M}(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}}} \ . \tag{29}$$

Furthermore, the scaling factor σ can be included in the update. We chose here a uniform distribution $\sigma \sim U(0.1, 10)$, leading to the derivative

$$\frac{\partial \log p(\boldsymbol{\theta}|\boldsymbol{D})}{\partial \sigma} = \frac{\partial p(\boldsymbol{D}|\boldsymbol{\theta})}{\partial \tilde{\sigma}} \cdot \frac{\partial \tilde{\sigma}}{\partial \sigma} - \sigma \tag{30}$$

$$= \frac{\partial p(\boldsymbol{D}|\boldsymbol{\theta})}{\partial \tilde{\sigma}} \cdot (10 - 0.1)\, \phi(\sigma) - \sigma \tag{31}$$

$$\frac{\partial p(\boldsymbol{D}|\boldsymbol{\theta})}{\partial \tilde{\sigma}} = -\frac{2m}{\tilde{\sigma}} - \frac{2}{\tilde{\sigma}^3} \sum_i^m (\boldsymbol{D}_i - \mathcal{M}(\boldsymbol{\theta}))^\top \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{D}_i - \mathcal{M}(\boldsymbol{\theta})) \ . \tag{32}$$

In addition to the variance diagnostic, we use the Bhattacharyya distance as a measure for the performance of TMCMC and TM approximations. The distance is calculated by

$$d_B(p_{ref}, p_{sim}) = -\log \int_X \sqrt{p_{ref}(x)\, p_{sim}(x)} \mathrm{d}x \tag{33}$$

where p_{ref} is the target posterior density and p_{sim} are the approximations. Since only samples are available for TMCMC, we use a Gaussian mixture model to obtain the PDF. As integration points we use samples obtained from a reference TMCMC run with 10^5 samples, assuming these cover the posterior density well enough. We then scale the resulting PDFs such that their integral over this domain is equal to unity:

$$p(x) = \frac{\hat{p}(x)}{\int_{X_{ref}} \hat{p}(x) \mathrm{d}x} \tag{34}$$

with X_{ref} as integration points from the reference TMCMC-run and $\hat{p}$ as and unscaled PDF. The integral in the denominator is calculated with MC-integration

$$\int_{X_{ref}} \hat{p}(x) \mathrm{d}x = \frac{1}{N_{ref}} \cdot \sum_i \hat{p}(x_i) \ . \tag{35}$$

The scaling is needed s.t. the PDFs become comparable since the unscaled form of the posterior is not available.

For both cases, we calculated three different setups for TMCMC (500, 1000, and 5000 samples per level) and five integration schemes for TM (MC, LHS, Gauss-Hermite, Smolyak, and sparse Smolyak). We also analyzed three different sample sizes for LHS sampling to make out differences that result from the number of integration points. Results are shown in Table 1. The number of integration points in the TM approach and the number of samples per level in TMCMC are denoted behind the method name in the table. Also shown is the total number of function calls for each setup.

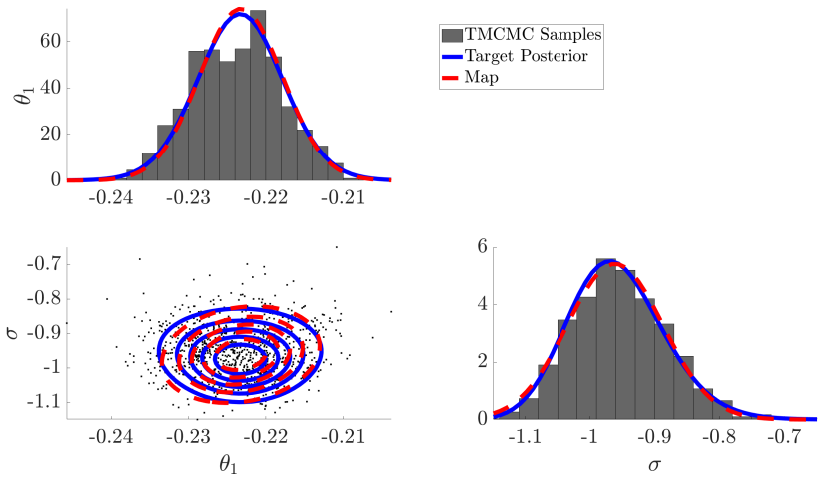
As is evident from the table, most of the TM approaches are more efficient than TMCMC in case I. TMCMC with 500 samples uses about 4.5 times more function evaluations than TM-Smolyak and 6.5 times more than TM-sparse Smolyak, while also being less accurate. With a higher number of samples, TMCMC becomes more accurate, however since it is a sampling algorithm the posterior approximation will be biased, as is shown by the lower accuracy in the TMCMC approaches. TM on the other hand directly approximates the posterior function and can thus achieve a higher accuracy, depending on the integration scheme. Using random grids such as MC or LHS seems to introduce some error to the TM approximation since both Smolyak integration schemes can approximate the posterior perfectly. Figure 5 illustrates this behavior,

$\boldsymbol{\theta}$	Method	function calls	ε_σ	d_B
Case I	TMCMC_500	4,000	-	0.00190
	TMCMC_1000	8,000	-	0.000949
	TMCMC_5000	40,000	-	0.000648
$\begin{bmatrix} k_1 \\ \sigma \end{bmatrix}$	TM_MC_20	1,156	0.0522	0.00837
	TM_LHS_10	505	0.540	0.0335
	TM_LHS_20	1,365	0.0255	0.00311
	TM_LHS_50	2,275	0.0159	0.00230
	TM_GaussHerm_16	665	0.432	0.0191
	TM_Smolyak_25	900	0.000462	0.0000496
	TM_SparseSmolyak_17	620	0.000505	0.0000496

$\boldsymbol{\theta}$	Method	function calls	ε_σ	d_B
Case II	TMCMC_500	6,500	-	0.628
	TMCMC_1000	13,000	-	0.137
	TMCMC_5000	65,000	-	0.00288
$\begin{bmatrix} k_1 \\ k_2 \\ k_3 \\ \sigma \end{bmatrix}$	TM_MC_50	5,000	0.286	0.226
	TM_LHS_20	2,110	0.180	0.223
	TM_LHS_50	4,800	0.0939	0.217
	TM_GaussHerm_256	15,666	0.738	0.248
	TM_Smolyak_87	4,748	0.00215	0.211
	TM_SparseSmolyak_49	2,696	0.00219	0.211

Table 1: Results of updating for Case I and II for the dynamical system with different approaches for TM and TMCMC. The number behind the method indicates the number of samples per level in TMCMC and the number of integration points in TM.

(a) LHS



(b) sparse Smolyak

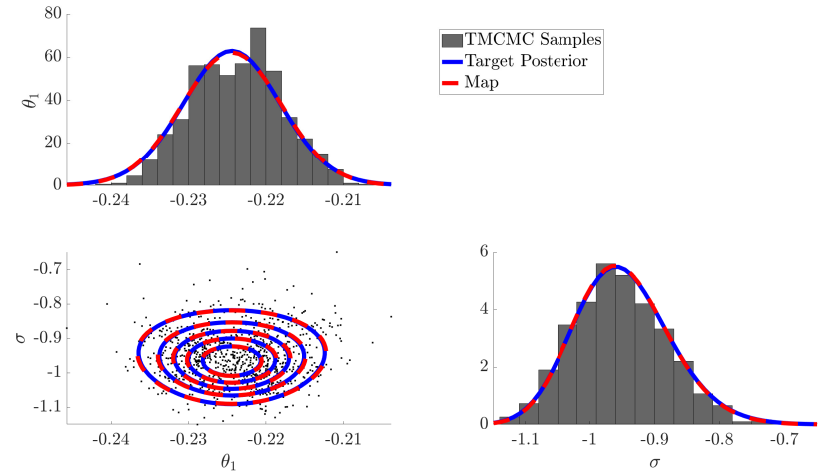


Figure 5: Results for calculation of maps obtained with (a) LHS and (b) sparse Smolyak integration, as well as samples from TMCMC, in case I for the dynamical system. Total function evaluations: LHS: 1365, sparse Smolyak: 620, TMCMC: 8000

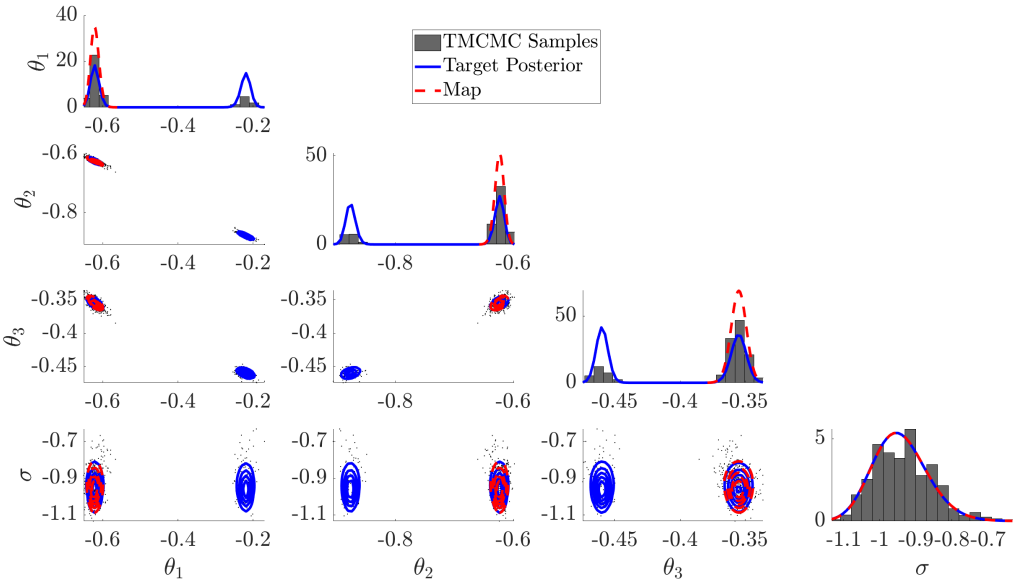


Figure 6: Comparison of target and approximated posterior PDFs $p(\boldsymbol{\theta}_{i,j}|\boldsymbol{D})$, TMCMC model evaluations: 13,000, sparse Smolyak model evaluation: 2,893

the final approximation with LHS shows some deviation from the true posterior, while sparse Smolyak integration fits the approximation perfectly. Due to some lumping of samples, MC and LHS sampling schemes possibly introduce some bias in some regions of the reference density, hence the approximation will also be biased in these regions. Especially while using only a few integration points in LHS, this behavior became more obvious. The difference in results between 10 and 50 integration points shows this as well.

For case II the posterior density becomes bi-modal. This poses multiple issues for both methods, however, due to the aforementioned annealing process, TMCMC was able to capture both modes in most settings. One issue for TM lies in the direction of the used KL divergence. Swapping the arguments in Eq. (10) could theoretically lead to a better performance. This would however require knowledge about the posterior's topology since the expectancy operator in Eq. (11) is then formulated with respect to the target density. Additionally, using a single Laplace approximation as first step limits the considered support in the posterior for $\boldsymbol{\theta}$. Furthermore, the posterior has to be absolutely continuous for the TM framework [36]. Posterior values of zero or very close to zero between the modes lead to plateaus which violate this criterion. This is a risk in applications like SHM where the posterior has a much smaller support than the prior. Because only one mode is captured, d_B in all cases for the TM-approach is high compared to the TMCMC approaches, however, as is illustrated in Fig. 6 the TM approach again works very well in capturing the posterior shape for one mode. Also shown is the fact that TMCMC is not able to capture well the ratios of both modes, as they would be nearly equally weighted, however TMCMC introduces some bias towards one of the peaks. All of the TM approaches again use less function evaluations than TMCMC-500, while also being more accurate. It was noticed that the low number of samples in TMCMC-500 meant that often only one mode was captured in the approximation. Smolyak and sparse Smolyak are again the most accurate for the TM

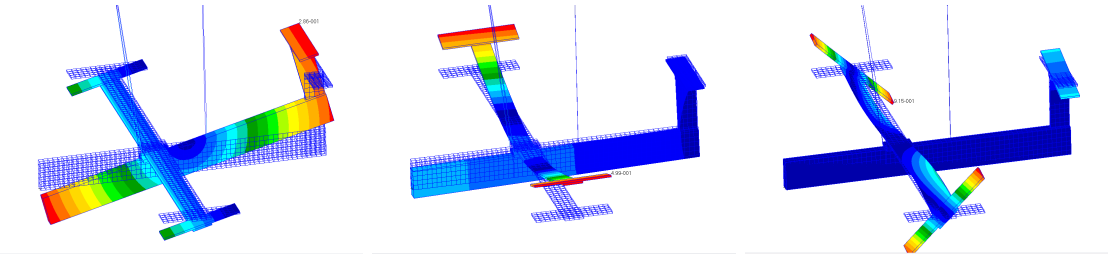


Figure 7: Illustration of the DLRAirmod example, screenshots taken from the FE software NASTRAN. The images show examples of computed mode shapes.

approaches, but because that one mode of the posterior is disregarded, they lack the accuracy of TMCMC with more samples. Note that the variance diagnostic in case II has comparably low values to case I, even though half of the posterior PDF was not captured. The variance diagnostic captures only the relationship between the calculated map and the reference density, it is therefore blind to the second mode and calculates low values. Due to the sampling nature and the slow convergence to the posterior, TMCMC can handle the multi-modality better.

4.2. DLR-Airmod

Here we investigate parameter updating on a higher dimensional, non-linear model for an experimental setup of an airplane structure (DLR-Airmod, for further details see [48, 49, 50], an illustration is shown in Figure 7). The structure is a replica of the GARTEUR SM-AG19 test bed, which was designed to investigate experimental determination of modal characteristics. The DLR-Airmod structure consists of an aluminum frame with several attached sensors that change the dynamic behavior, so the original goal was to update an FE-model of the structure to simulate this influence based on vibration measurements. We concentrate in this example on 11 input parameters to the model plus the scaling parameter σ, giving a total of 12 parameters to update. The 11 model inputs denote mass and stiffness parameters of structural parts with attached sensors and cables. The model provides 10 output parameters, which correspond to modal frequencies. A list of the considered input and output parameters with their physical mean values is shown in Table 2. Instead of the full FE-model, we used a neural network as a surrogate model, which was trained on 10^5 samples from the original FE-model of the system, where the input consists of evenly spaced samples in a range of 0.1 to 2 times the mean value for each parameter. To normalize the training data it is ideal to keep this setup, s.t. the parameters have equal order of magnitude. As a prior distribution we therefore chose a uniform distribution $\tilde{\theta}_i \sim \mathrm{U}(0.1, 2)$, where $\tilde{\theta}_i$ is the i-th input in physical space. Note that the physical space in this example denotes the normalized inputs. Moreover, the inputs are again mapped to standard normal space using Eq. (25) with a lower bound of 0.1 and an upper bound of 2. Using a neural network as a surrogate model comes with the advantage of being differentiable and thus usable in conjunction with the presented setup of transport maps. The data is generated from the numerical model using a nominal value of 0.3 for all entries of $\boldsymbol{\theta}$. However, to find out about differences in the approaches for different support sizes of the posterior we look at two different cases, one where the variance in the input is 10 % and the other where the variance is 1 % of the mean value. The posterior is of the same form as Eqs. (26) and (30) with $m = 50$, $|\boldsymbol{D}| = 10$ and $\sigma \sim \mathrm{U}(1, 50)$. $\mathcal{M}(\boldsymbol{\theta})$ is here the output of the neural network, hence $\partial\mathcal{M}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be calculated with automatic differentiation.

Inputs		
Parameter	Description	Mean physical value
θ_1	Front bungee cord stiffness	$3.375\,\mathrm{N/m^2}$
θ_2	Rear bungee cord stiffness	$14.060\,\mathrm{N/m^2}$
θ_3	Vertical tail plane stiffness	$243.75\,\mathrm{N/m}$
θ_4	Tail joint mass	$0.375\,\mathrm{kg}$
θ_5	Right wing-tip mass	$0.345\,\mathrm{kg}$
θ_6	Left wing-tip mass	$0.345\,\mathrm{kg}$
θ_7	Right/Left wing-tip additional mass	$0.028\,\mathrm{kg}$
θ_8	Outer wing additional mass	$0.028\,\mathrm{kg}$
θ_9	Wing-fuselage joint stiffness, x-direction	$3.75 \times 10^7\,\mathrm{N/m}$
θ_{10}	Wing-fuselage joint stiffness, y-direction	$3.75 \times 10^7\,\mathrm{N/m}$
θ_{11}	Wing-fuselage joint stiffness, z-direction	$1.31 \times 10^7\,\mathrm{N/m}$

Outputs		
Parameter	Description	Mean physical value
D_1	RBM roll	$0.697\,\mathrm{Hz}$
D_2	RBM pitch	$1.04\,\mathrm{Hz}$
D_3	RBM heave	$2.81\,\mathrm{Hz}$
D_4	2nd EF wing bending	$5.57\,\mathrm{Hz}$
D_5	3rd EF wing bending	$14.69\,\mathrm{Hz}$
D_6	antisymm. wing torsion	$29.33\,\mathrm{Hz}$
D_7	symm. wing torsion	$31.96\,\mathrm{Hz}$
D_8	vert. tail plane bending	$35.07\,\mathrm{Hz}$
D_9	4th EF wing bending	$44.08\,\mathrm{Hz}$
D_{10}	1st EF wing foreaft	$47.13\,\mathrm{Hz}$

Table 2: Inputs and outputs for the DLRAirmod model. **RBM**: rigid body motion, **EF**: eigenfrequency. Data obtained from [47].

Due to the findings in the previous section, we used TMCMC with 1000 and 5000 samples per level and TM with LHS (100 integration points) and sparse Smolyak (337 integration points) integration schemes. Furthermore, for the initial Laplace map in the TM approaches we increased the robustness by using a global search approach to find $\boldsymbol{\theta}_0$. We found that using eight randomly sampled starting points instead of starting at zero for the optimization was sufficient to not converge to a local minimum. While this increases the number of needed model evaluations, it also makes the TM approaches more robust since the calculated maps are conditioned on the global instead of a local minimum. A selection of resulting density approximations in the case of wide support can be found in Figure 8 and for the small support in Figure 9. Because the integration of the true posterior in 12 dimensions was not possible anymore, the densities shown are conditioned on the known posterior mode $\boldsymbol{\theta}_0 = [0.3, \ldots, 0.3, -0.6]$. The numbers of function evaluations are shown in Table 3. The estimated mean values for $\boldsymbol{\theta}$ for the 10 % variance case are shown in Table 4 and for the 1 % variance case in Table 5. These results show an overall good agreement, as most values are close to the prescribed mean of 0.3 for $\theta_{1,\ldots,11}$. For TMCMC, some of the values are quite far from the true value, indicating convergence to a local posterior maximum. In general, due to the high dimensionality of the posterior, this problem is difficult to solve for any approach, which is further illustrated in the model outputs shown in Figure 10. Even though the approximations show variations, when samples obtained from both approaches are propagated through the model, output and data are overlapping closely.

In the case of the wider support (Figure 8), there are only small differences between the two TM approaches, while LHS integration uses only about 75% of the model evaluations of the sparse Smolyak approach. Between using 1000 and 5000 samples per level in TMCMC, there is a larger difference in the approximation quality, but also a sixfold increase in function evaluations. Even though the posterior is slightly multi-modal and thus the approximations become even more difficult, generally it can be seen that the TM approaches capture the posterior better than TMCMC. This is also evidenced by the estimated mean values (Table 4). Especially for θ_6 to θ_{10} the estimations are relatively far away from the true value. Since more integration points are needed for the TMs, the difference in function evaluations to TMCMC becomes less.

The narrower support case (Figure 9) is more difficult to approximate, hence both methods struggle to capture the posterior well. Again, TMCMC-1000 has the lowest cost with the lowest performance and TMCMC-5000 has the highest cost, although it does not perform as well as the TM-approximations in most dimensions. TM uses about 45% more function evaluations in the small support case when compared to the wide support, the increase for TMCMC is about 30%. The largest deviations in the mean value estimations (Table 5) are observed again in parameters θ_6 to θ_{10}. In this case the TM approach deviates more than in the previous example, but not as much as TMCMC.
These results show an advantage of TM approaches, the posterior function is directly approximated as a whole and thus can be captured more accurately. TMCMC (and sampling methods in general) does not have this kind of global information and is therefore prone to converge to local minima. One way of avoiding this is to increase the number of samples, which in turn also increases the cost.

5. Conclusion

A comparison between a Markov Chain Monte Carlo (namely Transitional MCMC) approach and a transport map approach on Bayesian posterior estimation was conducted. Both algorithms

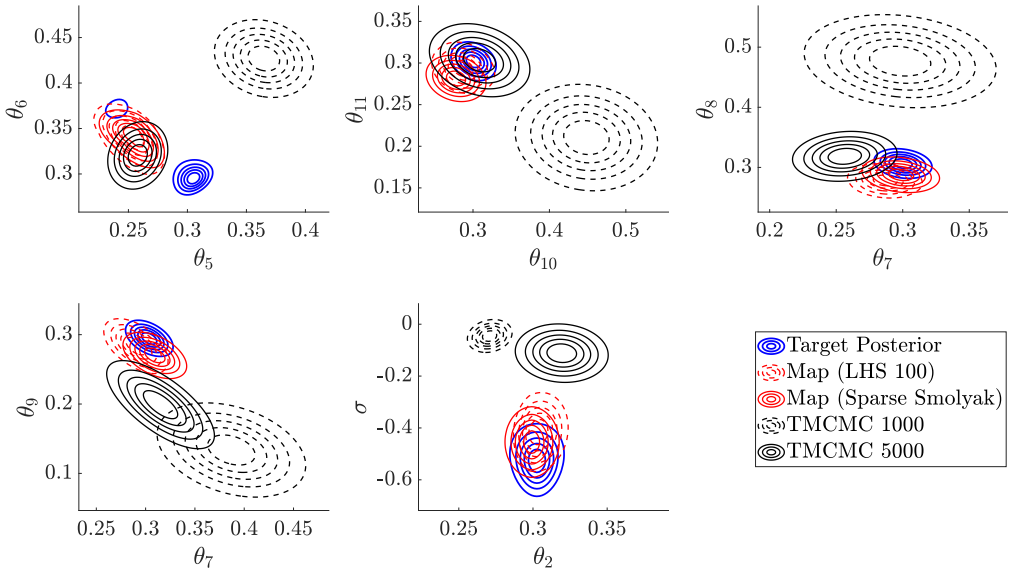


Figure 8: Selected approximation results for conditional PDF $p(\boldsymbol{\theta}_{i,j}|\boldsymbol{D}, \boldsymbol{\theta}_0)$ with $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{11}, \sigma]$ for the case of 10 % variance in the DLR-Airmod-model with various approaches. PDFs are conditioned on the known posterior mode $\boldsymbol{\theta}_0 = [0.3, \ldots, 0.3, -0.6]^\top$, PDF for TMCMC obtained with Gaussian mixture model.

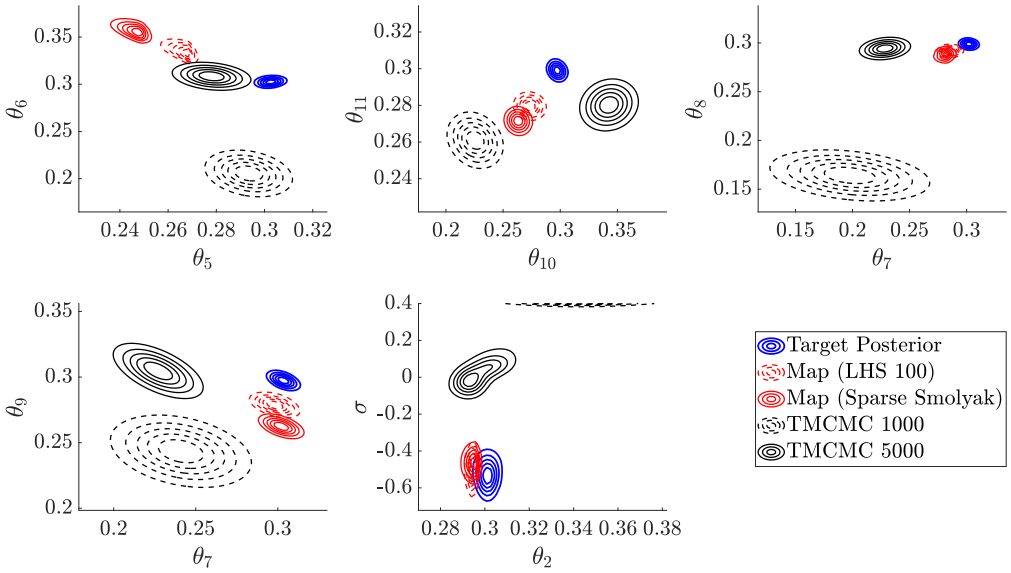


Figure 9: Selected approximation results for conditional PDF $p(\boldsymbol{\theta}_{i,j}|\boldsymbol{D}, \boldsymbol{\theta}_0)$ with $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{11}, \sigma]$ for the case of 1 % variance in the DLR-Airmod-model with various approaches. PDFs are conditioned on the known posterior mode $\boldsymbol{\theta}_0 = [0.3, \ldots, 0.3, -0.6]^\top$, PDF for TMCMC obtained with Gaussian mixture model.

Method	Large support	Small support
TMCMC_1000	16,000	21,000
TMCMC_5000	95,000	120,000
TM_LHS	30,427	43,427
TM_SparseSmolyak	40,476	60,013

Table 3: Number of function evaluations for updating of DLR-Airmod model for large support and small support case

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
TM	0.297	0.308	0.288	0.300	0.233	0.371	0.301	0.306
TMCMC	0.283	0.294	0.305	0.284	0.379	0.513	-0.042	0.510
	θ_9	θ_{10}	θ_{11}	θ_{12}				
TM	0.310	0.305	0.315	-0.545				
TMCMC	0.684	-0.035	0.290	0.015				

Table 4: Estimated posterior mean values of TMCMC with 5000 samples and TM with sparse Smolyak integration, 10 % variance

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
TM	0.300	0.306	0.304	0.304	0.232	0.354	0.341	0.278
TMCMC	0.310	0.285	0.308	0.303	0.297	0.471	0.062	0.428
	θ_9	θ_{10}	θ_{11}	θ_{12}				
TM	0.275	0.343	0.314	-0.546				
TMCMC	0.541	0.074	0.325	0.153				

Table 5: Estimated posterior mean values of TMCMC with 5000 samples and TM with sparse Smolyak integration, 1 % variance

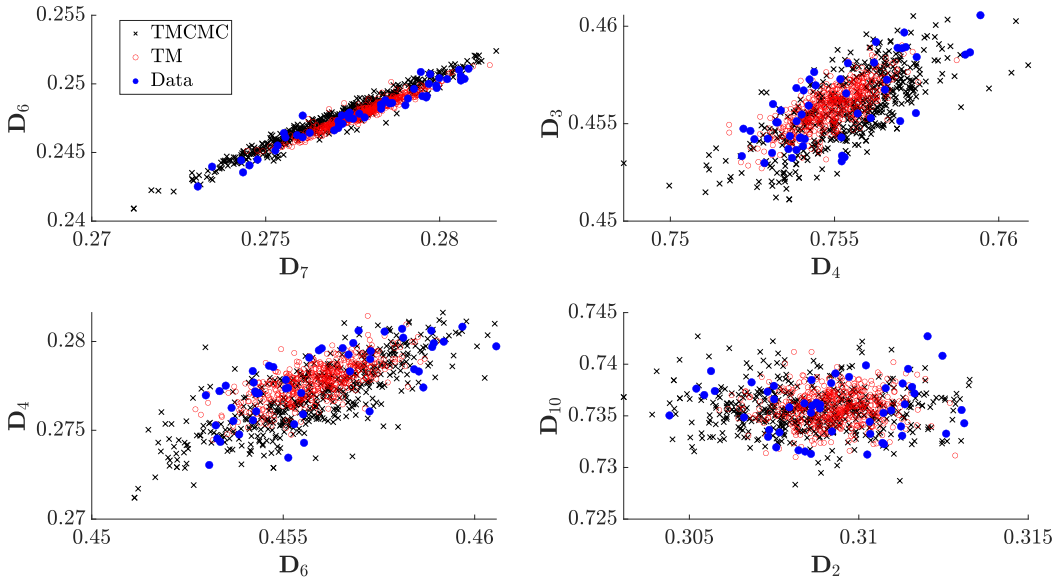


Figure 10: Selected outputs of the DLR-Airmod-model for samples from TMCMC-5000 and TM-sparse Smolyak approaches, output for the 1 %-variance case

were applied to two dynamic models with varying complexity to highlight differences between the two approaches. It was found that overall, transport map approaches are more efficient and accurate than TMCMC, although they are in the setup used in this work not able to capture multi-modal posterior functions. There are several differences in the two approaches with TMCMC being generally easier to implement but also showing a larger degree of variation in the results. For MCMC methods in general, the number of samples and Markov chains needs to be set a-priori. Often it is not clear how large these numbers should be, although they have a significant influence on the accuracy. Too few samples introduce a bias in the posterior, while too many samples give redundant information and therefore waste computation time. Many MCMC algorithms also require burn-in samples, which result in additional computation costs. Through the variational formulation, transport maps can use global information about the approximation quality and thus are more accurate. To solve the optimization procedure for TM, gradients are useful, however, these might not be available in some situations. While the setups used in this paper are easily differentiable, it remains an open question if gradient-free optimization procedures are efficient, especially in higher dimensional settings. To overcome this problem, surrogate systems or model order reduction methods can be applied. There also exist MCMC schemes using gradients in which cases using the TM framework can be an option. TMs provide an analytic function of the posterior, which allows for probability computations that are not easily available in sampling approaches. In the latter cases, it is common to fit a probability distribution to the samples, which adds a layer of approximation and thus decreases the accuracy. Transport maps are able to directly provide this relationship and are less restrictive to the shape of the posterior, as technically any continuous PDF can be approximated. This also allows the drawing of samples directly from the posterior, whereas with sampling methods the model is needed to draw new samples. Addressing multi-modal posterior distributions with the TM

framework requires more work. In the used setup it was difficult to capture multiple modes once the approximation converged to a single mode. This is a challenge for most posterior estimation methods and could be overcome by using annealing or sequential MC in conjunction with TM. It could also possible to use multiple Laplace approximations and calculate independent maps by splitting the posterior space into multiple regions. However, arising issues with weighting and discontinuities between the regions need to be addressed.

As was shown in the numerical examples in this paper, while transport maps still outperform TMCMC in higher-dimensional problems, the efficiency gain is not as high anymore. This problem can be reduced (for both TM and MCMC approaches) by using dimensionality reduction methods on the posterior density which exploit the fact that the data is generally not informative in the full parameter space [51, 52]. This is also possible in an a-priori step before observing data, which is useful for sequential updating [53]. However, further research is needed to incorporate these techniques into structural health monitoring problems.

Acknowledgment

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the GRK2657 (grant reference number 433082294) is greatly appreciated.

Author contributions

J. Grashorn: Conceptualization, Methodology, Software, Writing- Original draft preparation, Visualization; **M. Broggi:** Conceptualization, Methodology, Validation, Writing- Reviewing and Editing; **L. Chamoin:** Conceptualization, Methodology, Writing- Reviewing and Editing, Supervision, Funding acquisition; **M. Beer:** Conceptualization, Writing- Reviewing and Editing, Supervision, Funding acquisition

References

[1] James L Beck, Lambros S Katafygiotis, et al. Updating models and their uncertainties. i: Bayesian statistical framework. *Journal of Engineering Mechanics-Proceedings of the ASCE*, 124(4):455–462, 1998.

[2] James L Beck. Bayesian system identification based on probability logic. *Structural Control and Health Monitoring*, 17(7):825–847, 2010.

[3] Antonios Kamariotis, Eleni Chatzi, and Daniel Straub. Value of information from vibration-based structural health monitoring extracted via bayesian model updating. *Mechanical Systems and Signal Processing*, 166:108465, 2022.

[4] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[5] Jinwoo Jang and Andrew Smyth. Bayesian model updating of a full-scale finite element model with sensitivity-based clustering. *Structural Control and Health Monitoring*, 24(11):e2004, 2017.

[6] James L Beck and Siu-Kui Au. Bayesian updating of structural models and reliability using markov chain monte carlo simulation. *Journal of engineering mechanics*, 128(4):380–391, 2002.

[7] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[8] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970.

[9] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

[10] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. Dram: efficient adaptive mcmc. *Statistics and computing*, 16:339–354, 2006.

[11] Daniel Straub and Iason Papaioannou. Bayesian updating with structural reliability methods. *Journal of Engineering Mechanics*, 141(3):04014134, 2015.

[12] FA DiazDelaO, A Garbuno-Inigo, SK Au, and I Yoshida. Bayesian updating and model class selection with subset simulation. *Computer Methods in Applied Mechanics and Engineering*, 317:1102–1121, 2017.

[13] Adolphus Lye, Alice Cicirello, and Edoardo Patelli. Sampling methods for solving Bayesian model updating problems: A tutorial. *Mechanical Systems and Signal Processing*, 159:107760, October 2021.

[14] Jianye Ching and Yi-Chu Chen. Transitional markov chain monte carlo method for bayesian model updating, model class selection, and model averaging. *Journal of engineering mechanics*, 133(7):816–832, 2007.

[15] Panagiotis Angelikopoulos, Costas Papadimitriou, and Petros Koumoutsakos. X-tmcmc: Adaptive kriging for bayesian inverse modeling. *Computer Methods in Applied Mechanics and Engineering*, 289:409–428, 2015.

[16] Wolfgang Betz, Iason Papaioannou, and Daniel Straub. Transitional markov chain monte carlo: observations and improvements. *Journal of Engineering Mechanics*, 142(5):04016016, 2016.

[17] Ka-Veng Yuen, James L Beck, and Siu Kui Au. Structural damage detection and assessment by adaptive markov chain monte carlo simulation. *Structural Control and Health Monitoring*, 11(4):327–347, 2004.

[18] Eloi Figueiredo, Lucian Radu, Keith Worden, and Charles R Farrar. A bayesian approach based on a markov-chain monte carlo method for damage detection under unknown sources of variability. *Engineering Structures*, 80:1–10, 2014.

[19] Heung-Fai Lam, Jiahua Yang, and Siu-Kui Au. Bayesian model updating of a coupled-slab system using field test data utilizing an enhanced markov chain monte carlo simulation algorithm. *Engineering Structures*, 102:144–155, 2015.

[20] Jianye Ching and Jiun-Shiang Wang. Application of the transitional markov chain monte carlo algorithm to probabilistic site characterization. *Engineering Geology*, 203:151–167, 2016.

[21] Yin-Fu Jin, Zhen-Yu Yin, Wan-Huan Zhou, and Suksun Horpibulsuk. Identifying parameters of advanced soil models using an enhanced transitional markov chain monte carlo method. *Acta Geotechnica*, 14:1925–1947, 2019.

[22] Xin Zhou, Chul-Woo Kim, Feng-Liang Zhang, and Kai-Chun Chang. Vibration-based bayesian model updating of an actual steel truss bridge subjected to incremental damage. *Engineering Structures*, 260:114226, 2022.

[23] Wang-Ji Yan, Dimitrios Chronopoulos, Costas Papadimitriou, Sergio Cantero-Chinchilla, and Guo-Shu Zhu. Bayesian inference for damage identification based on analytical probabilistic model of scattering coefficient estimators and ultrafast wave scattering simulation scheme. *Journal of Sound and Vibration*, 468:115083, 2020.

[24] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[25] Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[26] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. Sampling via measure transport: An introduction. In *Handbook of Uncertainty Quantification*, pages 1–41. Springer International Publishing, 2016.

[27] Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[28] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[29] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[30] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[31] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[32] Tiangang Cui and Sergey Dolgov. Deep composition of tensor-trains using squared inverse rosenblatt transports. *Foundations of Computational Mathematics*, 22(6):1863–1922, 2022.

[33] Sergey Dolgov, Karim Anaya-Izquierdo, Colin Fox, and Robert Scheichl. Approximation and sampling of multivariate probability distributions in the tensor train decomposition. *Statistics and Computing*, 30:603–625, 2020.

[34] Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. On the representation and learning of monotone triangular transport maps. *Foundations of Computational Mathematics*, pages 1–46, 2023.

[35] Jan Grashorn. Dataset: Transport map bayesian updating for shm, 2023. `https://doi.org/10.25835/nkmp900p`.

[36] Matthew D Parno and Youssef M Marzouk. Transport map accelerated markov chain monte carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.

[37] MParT Development Team. Monotone Parameterization Toolkit (MParT), 2022. Version 1.2.0.

[38] Paul-Baptiste Rubio, François Louf, and Ludovic Chamoin. Transport map sampling with pgd model reduction for fast dynamical bayesian data assimilation. *International Journal for Numerical Methods in Engineering*, 120(4):447–472, 2019.

[39] Paul-Baptiste Rubio, Ludovic Chamoin, and François Louf. Real-time bayesian data assimilation with data selection, correction of model bias, and on-the-fly uncertainty propagation. *Comptes Rendus Mécanique*, 347(11):762–779, 2019.

[40] Max Gunzburger, CG Webster, and Guannan Zhang. Sparse collocation methods for stochastic interpolation and quadrature. *Handbook of uncertainty quantification*, pages 1–46, 2017.

[41] C. Piazzola and L. Tamellini. The Sparse Grids Matlab kit - a Matlab implementation of sparse grids for high-dimensional function approximation and uncertainty quantification. *ArXiv*, (2203.09314), 2023.

[42] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

[43] Michael Arbel, Alex Matthews, and Arnaud Doucet. Annealed flow transport monte carlo. In *International Conference on Machine Learning*, pages 318–330. PMLR, 2021.

[44] Tiangang Cui, Sergey Dolgov, and Olivier Zahm. Scalable conditional deep inverse rosenblatt transports using tensor trains and gradient-based dimension reduction. *Journal of Computational Physics*, 485:112103, 2023.

[45] Tiangang Cui, Sergey Dolgov, and Olivier Zahm. Self-reinforced polynomial approximation methods for concentrated probability densities. *arXiv preprint arXiv:2303.02554*, 2023.

[46] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[47] Yves Govers, H Haddad Khodaparast, M Link, and JE Mottershead. A comparison of two stochastic model updating methods using the dlr airmod test structure. *Mechanical Systems and Signal Processing*, 52:105–114, 2015.

[48] Yves Govers and Michael Link. Stochastic model updating of an aircraft like structure by parameter covariance matrix adjustment. In *Proc. of the International Conference on Noise and Vibration Engineering, ISMA*, 2010.

[49] Edoardo Patelli, Matteo Broggi, Yves Govers, and John E Mottershead. Model updating strategy of the dlr-airmod test structure. *Procedia Engineering*, 199:978–983, 2017.

[50] Edoardo Patelli, Yves Govers, Matteo Broggi, Herbert Martins Gomes, Michael Link, and John E Mottershead. Sensitivity or bayesian model updating: a comparison of techniques using the dlr airmod test data. *Archive of Applied Mechanics*, 87:905–925, 2017.

[51] Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, and Youssef Marzouk. Greedy inference with structure-exploiting lazy maps. *Advances in Neural Information Processing Systems*, 33:8330–8342, 2020.

[52] Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, and Youssef Marzouk. Certified dimension reduction in nonlinear Bayesian inverse problems. *Mathematics of Computation*, 91(336):1789–1835, April 2022.

[53] Tiangang Cui and Olivier Zahm. Data-free likelihood-informed dimension reduction of Bayesian inverse problems. *Inverse Problems*, 37(4):045009, April 2021.