Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Importance Sampling-Based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models

Kjartan Kloster Osmundsen, Tore Selland Kleppe & Roman Liesenfeld

To cite this article: Kjartan Kloster Osmundsen, Tore Selland Kleppe & Roman Liesenfeld (2021): Importance Sampling-Based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2021.1923519

To link to this article: https://doi.org/10.1080/10618600.2021.1923519

View supplementary material

Published online: 18 Jun 2021.

Submit your article to this journal

Article views: 68

View related articles

View Crossmark data

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/10618600.2021.1923519

Taylor & Francis
Taylor & Francis Group

Importance Sampling-Based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models

Kjartan Kloster Osmundsen[a], Tore Selland Kleppe[a], and Roman Liesenfeld[b]

[a]Department of Mathematics and Physics, University of Stavanger, Stavanger, Norway; [b]Institute of Econometrics and Statistics, University of Cologne, Cologne, Germany

ABSTRACT

We propose an importance sampling (IS)-based transport map Hamiltonian Monte Carlo procedure for performing a Bayesian analysis in nonlinear high-dimensional hierarchical models. Using IS techniques to construct a transport map, the proposed method transforms the typically highly complex posterior distribution of a hierarchical model such that it can be easily sampled using standard Hamiltonian Monte Carlo. In contrast to standard applications of high-dimensional IS, our approach does not require IS distributions with high fidelity, which makes it computationally very cheap. Moreover, it is less prone to the notorious problem of IS that the variance of IS weights can become infinite. We illustrate our algorithm with applications to challenging dynamic state-space models, where it exhibits very high simulation efficiency compared to relevant benchmarks, even for variants of the proposed method implemented using a few dozen lines of code in the Stan statistical software. The article is accompanied by supplementary material containing further details, and the computer code is available at https://github.com/kjartako/TMHMC. These are also supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2020
Revised March 2021

KEYWORDS
Dynamic rescaling; Latent variables; MCMC; Stan; State-space model

1. Introduction

Computational methods for high-dimensional Bayesian nonlinear/non-Gaussian hierarchical models is an active field of research, and advances in such methods allow researchers to build and analyze progressively more complex models. Existing Markov chain Monte Carlo (MCMC) methods for such models can broadly classified into four categories. The first category consists of Gibbs sampling procedures which are widely used in part because they are easy to implement (Robert and Casella 2004). However, a naive implementation updating latent variables in one block and model parameters in another block can suffer from a very slow exploration of the target distribution if this joint distribution implies a strong dependence between the variables in the two blocks (Jacquier, Polson, and Rossi 1994). The second category includes methods that jointly update latent variables and parameters and thus avoid the dependence problem of Gibbs sampling. One such approach is to use Riemann manifold Hamiltonian Monte Carlo (RMHMC) methods (Girolami and Calderhead 2011; Zhang and Sutton 2014; Kleppe 2018). However, they critically require proposals which are properly aligned with the (typically fairly variable) local geometry of the target, the generation of which can be computationally demanding for complex high-dimensional joint posteriors of the parameters and latent variables. The third category is pseudo-marginal methods which bypass the dependence problem of Gibbs sampling by targeting directly the marginal posterior of theparameters (Andrieu, Doucet, and Holenstein 2010; Pitt et al. 2012). However, they require low variance, unbiased Monte Carlo (MC) estimates of that marginal posterior, which can often be computationally extremely demanding for high-dimensional models (Flury and Shephard 2011). In addition, for models with many parameters, it can be difficult to select an efficient proposal distribution for updating the parameters, especially if the MC estimates for the marginal posterior are noisy and/or contain many discontinuities, which is typically the case if the MC estimator is implemented using particle filtering techniques.

Finally, the fourth category is transport map/dynamic rescaling methods (Parno and Marzouk 2018; Hoffman et al. 2019). They modify the original (model implied) parameterization by using a nonlinear transport map (TM). The TM is chosen so that the target distribution in the modified parameterization is better behaved and allows MCMC sampling using standard techniques. For a specific class of nonlinear hierarchical models satisfying fairly restrictive regularity conditions, the dynamically rescaled Hamiltonian Monte Carlo (DRHMC) approach of Kleppe (2019) provided a recipe for constructing such TMs.

In this article, we also adopt a TM approach for high-dimensional Bayesian hierarchical models which enables HMC sampling from the joint posterior of the parameters and latent variables. Specifically, we propose to use TMs for the latent variables resulting from well-known importance sampling

CONTACT Tore Kleppe tore.kleppe@uis.no Department of Mathematics and Physics, University of Stavanger, 4036 Stavanger, Norway

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

(IS) methods. This IS-TM-HMC approach only requires that the joint posterior distribution is sufficiently smooth and can be evaluated up to a normalizing constant. Thereby it bypasses the significant requirements for analytical tractability that are needed for DRHMC. As a result, our approach is more automated and can in particular also be applied to a larger range of nonlinear models than DRHMC. Another advantageous feature of our IS-TM-HMC approach is that, unlike conventional high-dimensional IS applications (see, e.g., Koopman, Shephard, and Creal 2009), the IS densities used need not necessarily be of high fidelity for simulation efficiency as long as they reflect the location and scale of the conditional posterior distribution of the latent variables. Since our approach uses HMC simulation to update the parameters and latent variables simultaneously, it also avoids the slow exploration of the target, which is characteristic of Gibbs methods. Moreover, in contrast to RMHMC approaches, our TM method enables the use of standard HMC and in particular can be implemented with minimal effort using statistical software like Stan. This being said, in principle, other MCMC methods with a similar design as HMC (i.e., joint updates of parameters and latent variables and applicability in high-dimensional but close to Gaussian settings) may also be used for sampling the modified parameterization.

Our approach exploiting IS to construct TMs for standard HMC sampling on the joint space of the parameters and latent variables can be interpreted as a special case of the pseudo-marginal HMC method of Alenlöv, Doucet, and Lindsten (2016), which uses IS to marginalize the latent variables. This special case results if, in their approach, the IS simulation sample size is $n = 1$. However, Alenlöv, Doucet, and Lindsten (2016) considered IS estimators using the prior of the latent variables as IS density which ignores the information about the location and scale of the latent variables in the data likelihood. In applications to models with high-dimensional latent variables, those "brute force" IS estimators are known to suffer from a prohibitively large variance, even for a very large n (Danielsson 1994). Therefore, pseudo-marginal HMC based on these brute force IS estimators is in general not well suited for such models.

The article is organized as follows: In Section 2, we outline HMC and its application to hierarchical models. Section 3 introduces IS-based TMs and Section 4 discusses specific choices for such TMs. Simulation experiments that examine the tradeoff between fidelity and computational costs of the various TMs are provided in Section 5 for models with univariate latent state processes and in Section 6 for a model with a multivariate state process. Section 7 concludes with some discussion. Supplementary material provides additional details on the key algorithms, and the codes used for the computations are available at *https://github.com/kjartako/TMHMC*.

2. Background

In what follows, we use $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ to denote the probability density function of a $N(\boldsymbol{\mu}, \Sigma)$ random vector evaluated at $\mathbf{x}$, while $\nabla_{\mathbf{z}}$ and $\nabla^2_{\mathbf{z}}$ are used, respectively, for the gradient/Jacobian and Hessian operator with respect to the vector $\mathbf{z}$.

2.1. HMC

Over the past decade, HMC introduced by Duane et al. (1987) have been extensively used as a general-purpose MCMC method, often applied for simulating from posterior distributions arising in Bayesian models (Neal 2011). HMC produces MCMC chains by using the dynamics of a synthetic Hamiltonian system as a proposal mechanism. An easy to use HMC implementation which automatically tunes all tuning parameters is available in the popular Bayesian modeling software Stan (Stan Development Team 2019).

Suppose one seeks to sample from an analytically intractable distribution with a density $\pi(\mathbf{q})$, $\mathbf{q} \in \mathbb{R}^s$, and a density kernel $\tilde{\pi}(\mathbf{q}) \propto \pi(\mathbf{q})$, which can be pointwise evaluated. To this end, HMC takes the variable of interest $\mathbf{q}$ as the "position coordinate" of a Hamiltonian system, which is complemented by an (artificial) "momentum variable" $\mathbf{p} \in \mathbb{R}^s$. The corresponding Hamiltonian function specifying the total energy of the dynamical system is given by

$$H(\mathbf{q}, \mathbf{p}) = -\log \tilde{\pi}(\mathbf{q}) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}, \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{s\times s}$ is a symmetric, positive-definite "mass matrix" representing an HMC tuning parameter. For near-Gaussian target distributions, for instance, setting $\mathbf{M}$ close to the precision matrix of the target enables HMC to produce proposals that are close to independent of the current state (see Neal 2011, sec. 4.1 for details). The law of motions under the dynamic system specified by the Hamiltonian H is determined by Hamilton's equations given by

$$\begin{aligned} \frac{d}{dt}\mathbf{p}(t) &= -\nabla_{\mathbf{q}} H\left(\mathbf{q}(t), \mathbf{p}(t)\right) = \nabla_{\mathbf{q}} \log \tilde{\pi}(\mathbf{q}), \\ \frac{d}{dt}\mathbf{q}(t) &= \nabla_{\mathbf{p}} H\left(\mathbf{q}(t), \mathbf{p}(t)\right) = \mathbf{M}^{-1}\mathbf{p}. \end{aligned} \tag{2}$$

The dynamics associated with Hamilton's equations preserves both the Hamiltonian (i.e., $dH\left(\mathbf{q}(t), \mathbf{p}(t)\right)/dt = 0$) and the Boltzmann distribution $\pi(\mathbf{q}, \mathbf{p}) \propto \exp\{-H(\mathbf{q}, \mathbf{p})\} \propto \tilde{\pi}(\mathbf{q})\, \mathcal{N}(\mathbf{p}|\mathbf{0}_s, \mathbf{M})$, in the sense that if $[\mathbf{q}(t), \mathbf{p}(t)] \sim \pi(\mathbf{q}, \mathbf{p})$, then $[\mathbf{q}(t+\tau), \mathbf{p}(t+\tau)] \sim \pi(\mathbf{q}, \mathbf{p})$ for any (scalar) time increment τ. Based on the latter property, a valid MCMC scheme for generating $\{\mathbf{q}^{(k)}\}_k \sim \pi(\mathbf{q})$ is to alternate between the following two steps: (i) Sample a new momentum $\mathbf{p}^{(k)} \sim N(\mathbf{0}_s, \mathbf{M})$ from the $\mathbf{p}$-marginal of the Boltzmann distribution; and (ii) use the Hamilton's equations (2) to propagate $[\mathbf{q}(0), \mathbf{p}(0)] = [\mathbf{q}^{(k)}, \mathbf{p}^{(k)}]$ for some increment τ to obtain $[\mathbf{q}(\tau), \mathbf{p}(\tau)] = [\mathbf{q}^{(k+1)}, \mathbf{p}^*]$ and discard $\mathbf{p}^*$. However, for all but very simple scenarios (like those with a Gaussian target $\pi(\mathbf{q})$) the transition dynamics according to Equation (2) does not admit closed-form solution, in which case it is necessary to rely on numerical integrators for an approximative solution. Provided that the numerical integrator used for that purpose is symplectic, the numerical approximation error can be exactly corrected by introducing an accept-reject (AR) step, which uses the Hamiltonian to compare the total energy of the new proposal for the pair $(\mathbf{q}, \mathbf{p})$ with that of the old pair inherited from the previous MCMC step (Neal 2011). Accordingly, each HMC update step consists of the following individual steps:

- Refresh the momentum $\mathbf{p}^{(k)} \sim N(\mathbf{0}_s, \mathbf{M})$.
- Use L symplectic integrator steps with time-step size ε to approximate the dynamics (2) starting from $(\mathbf{q}(0), \mathbf{p}(0)) = (\mathbf{q}^{(k)}, \mathbf{p}^{(k)})$ and obtain $(\mathbf{q}^*, \mathbf{p}^*) \approx (\mathbf{q}(L\varepsilon), \mathbf{p}(L\varepsilon))$.
- Set $\mathbf{q}^{(k+1)} = \mathbf{q}^*$ with probability $\tilde{\alpha} = \min(1, \exp(H(\mathbf{q}^{(k)}, \mathbf{p}^{(k)}) - H(\mathbf{q}^*, \mathbf{p}^*))$ and $\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)}$ with probability $1 - \tilde{\alpha}$.

The most commonly used symplectic integrator is the Störmer-Verlet or leapfrog integrator (Leimkuhler and Reich 2004; Neal 2011). When implementing numerical integrators with AR-corrections it is critical that the selection of the step size accounts for the inherent tradeoff between the computing time required for generating AR proposals and their quality reflected by their corresponding acceptance rates. $(\mathbf{q}, \mathbf{p})$-proposals generated by using small (big) step sizes tend to be computationally expensive (cheap) but are typically numerically stable (unstable) and imply small (large) energy errors and thus high (low) acceptance rates (see, e.g., Leimkuhler and Reich 2004, chap. 2.6 for a discussion of stability). For a given step size, the numerical stability and the size of energy errors of symplectic integrators typically strongly depend not only on the dimension of the target distribution but also on its shape. Here, we adopt the general rule of thumb that high-dimensional targets with large deviations from a Gaussian shape typically require smaller step sizes and more steps for efficient simulation than high-dimensional near-Gaussian targets.

2.2. Hierarchical Models and HMC

Consider a stochastic model for a collection of observed data $\mathbf{y}$ involving a collection of latent variables $\mathbf{x} \in \mathbb{R}^D$ and a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ with prior density $p(\boldsymbol{\theta})$. The conditional likelihood for the observations $\mathbf{y}$ given a value of $\mathbf{x}$ is denoted by $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ and the prior for $\mathbf{x}$ by $p(\mathbf{x}|\boldsymbol{\theta})$. This latent variable model is assumed to be nonlinear and/or non-Gaussian so that both the joint posterior for $(\mathbf{x}, \boldsymbol{\theta})$ as well as the marginal posterior for $\boldsymbol{\theta}$ are analytically intractable. It is further assumed that $p(\boldsymbol{\theta})$, $p(\mathbf{x}|\boldsymbol{\theta})$ and $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ can be evaluated pointwise and have computable continuous derivatives in $(\mathbf{x}, \boldsymbol{\theta})$ up to order two.

The joint posterior for $(\mathbf{x}, \boldsymbol{\theta})$ under such a latent variable model, given by $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, can have a complex dependence structure. This is especially the case when the scale of $\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}$ varies substantially as a function of $\boldsymbol{\theta}$ in high-density regions of $p(\boldsymbol{\theta}|\mathbf{y})$, which leads to a "funnel-shaped" joint posterior (see Kleppe 2019, Fig. 1). In such cases, standard HMC for $\mathbf{q} = (\mathbf{x}^T, \boldsymbol{\theta}^T)^T$ must be tuned for the regions of the target distribution with the most extreme scale in order to ensure a numerically stable exploration of the full target distribution. This in turn leads to a computationally wasteful exploration of regions with a less extreme scale, since standard HMC rules out that tuning parameters adapt to the value of $\mathbf{q}$. Furthermore, the integrator step sizes (and mass matrices) found in an initial tuning phase depend crucially on the region with the most extreme scaling that was visited in this phase. If the regions where the target distribution has its most extreme scale are not visited during the tuning phase, then HMC may not explore them at all.

3. Transport Maps Based on IS Densities

In order to avoid the above-mentioned tuning problems of standard HMC, while bypassing computationally intensive $\mathbf{q}$-dependent tuning such as used by RMHMC, our proposed approach "preconditions" the original target by using TMs, so that the resulting modified target is close to Gaussian. This makes it suitable for statically tuned standard HMC. Such preconditioning, which aims at producing more tractable target distributions for MCMC methods has a long tradition, and prominent examples are the affine reparameterizations common for Gibbs sampling in regression models (Gelman et al. 2014, chap. 12). More recent such approaches are the semi-parametric TM procedure of Parno and Marzouk (2018) and the neural TM technique as described by Hoffman et al. (2019). Both approaches require that some samples from the original posterior are available for fitting or training (possibly several times) the TM. The TM approach followed here is based on analytical arguments rather than such posterior samples and therefore has similarities with the DRHMC method of Kleppe (2019). DRHMC constructs the TMs using a-priori knowledge about the precision and Fisher information matrices associated with the different conditional distributions of the model. However, the invertibility of the TMs in DRHMC necessitates that the model's conditional distributions have the so-called constant-information parameterization, which may be difficult to find for nonstandard distributions. Our strategy for constructing TMs exploits the IS principle, which does not require the availability of precision matrices or special parameterizations, and is therefore applicable to a larger class of nonlinear hierarchical models than DRHMC.

3.1. Transport Maps for Bayesian Hierarchical Models

A TM is a smooth bijective mapping Γ which relates the original parameterization $\mathbf{q} \sim \pi_{\mathbf{q}}(\mathbf{q})$ and a modified parameterization $\mathbf{q}'$ via $\mathbf{q} = \Gamma(\mathbf{q}')$. If $\mathbf{q}'$ is a random draw from the "pullback density" $\pi_{\mathbf{q}'}(\mathbf{q}') = \pi_{\mathbf{q}}(\Gamma(\mathbf{q}'))|\nabla_{\mathbf{q}'}\Gamma(\mathbf{q}')|$, then it can be transformed into a draw from $\pi_{\mathbf{q}}$ by simply applying the TM to $\mathbf{q}'$. The objective associated with the use of TMs is to select a map Γ so that the resulting $\pi_{\mathbf{q}'}$ is better suited for MCMC sampling than the original target $\pi_{\mathbf{q}}$. For an introduction of the concept of TMs to improve MCMC sampling efficiency, see Parno and Marzouk (2018, sec. 2). When applied to HMC sampling, this objective can be formulated as constructing the TM Γ such that $\pi_{\mathbf{q}'}$ comes as close as possible to a normal distribution with independent elements.

For our application to Bayesian hierarchical models, we consider TMs that are only non-trivial for the latent variables such that

$$\mathbf{q} = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{x} \end{bmatrix} = \Gamma(\mathbf{q}') = \begin{bmatrix} \boldsymbol{\theta} \\ \gamma_{\boldsymbol{\theta}}(\mathbf{u}) \end{bmatrix}, \qquad \mathbf{q}' = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{u} \end{bmatrix}.$$

The component of the TM specific to the latent variables, $\gamma_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^D$ is assumed to be a smooth bijective mapping for any admissible $\boldsymbol{\theta}$. We also assume that $\gamma_{\boldsymbol{\theta}}$ is smooth in $\boldsymbol{\theta}$ so that the complete map Γ is smooth and bijective. Since $\nabla_{\mathbf{u}}\boldsymbol{\theta} = \mathbf{0}$, it follows that $|\nabla_{\mathbf{q}'}\Gamma(\mathbf{q}')| = |\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})|$, and thus the modified target distribution (the pullback of $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ under Γ) has the form:

$$\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) \propto |\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| p(\boldsymbol{\theta}) \left[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})\right]_{\mathbf{x}=\gamma_{\boldsymbol{\theta}}(\mathbf{u})}. \quad (3)$$

Notice that the evaluation of Equation (3) for a specific $(\boldsymbol{\theta}, \mathbf{u})$-value requires to compute the corresponding value in the original parameterization of the latent variables $(\boldsymbol{\theta}, \mathbf{x}) = (\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}}(\mathbf{u}))$. Thus, MCMC simulations of $(\boldsymbol{\theta}, \mathbf{u})$ targeting (3) provide MCMC samples of $(\boldsymbol{\theta}, \mathbf{x})$ at no additional costs.

Now, let $m(\mathbf{x}|\boldsymbol{\theta})$ denote the "pushforward density" of a $N(\mathbf{0}_D, \mathbf{I}_D)$ for $\mathbf{u}$ under $\gamma_{\boldsymbol{\theta}}$, that is, the density of $\mathbf{x} = \gamma_{\boldsymbol{\theta}}(\mathbf{u})$ when $\mathbf{u} \sim N(\mathbf{0}_D, \mathbf{I}_D)$. Then the Jacobian determinant in Equation (3) can be expressed as $|\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| = \mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D)/\left[m(\mathbf{x}|\boldsymbol{\theta})\right]_{\mathbf{x}=\gamma_{\boldsymbol{\theta}}(\mathbf{u})}$, so that the modified target can be represented as follows:

$$\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) \propto \mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D)p(\boldsymbol{\theta})\omega_{\boldsymbol{\theta}}(\mathbf{u}),$$
$$\omega_{\boldsymbol{\theta}}(\mathbf{u}) = \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})}{m(\mathbf{x}|\boldsymbol{\theta})}\right]_{\mathbf{x}=\gamma_{\boldsymbol{\theta}}(\mathbf{u})}. \quad (4)$$

Representation (4) reveals that $\omega_{\boldsymbol{\theta}}(\mathbf{u}) = p(\mathbf{y}|\boldsymbol{\theta}) \;\forall\; \mathbf{u}$ if the pushforward density $m(\mathbf{x}|\boldsymbol{\theta})$ as a function in $\mathbf{x}$ is equal to the conditional posterior of the latent variables $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$. In this case, the modified target would be reduced to $\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) \propto \mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D)p(\boldsymbol{\theta}|\mathbf{y})$, so that the parameters and latent variables would be completely "decoupled" (see also Alenlöv, Doucet, and Lindsten 2016, for a similar discussion). Provided that the posterior of the parameters $p(\boldsymbol{\theta}|\mathbf{y})$ is reasonably well-behaved, such an "ideally" modified target would be well suited for HMC sampling. Of course, such an ideal modification of the target is infeasible when the model under consideration is nonlinear/non-Gaussian since neither $p(\boldsymbol{\theta}|\mathbf{y})$ nor $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ will have an analytically tractable form. However, this motivates our proposed approach to achieve a high HMC sampling efficiency. It consists in choosing the TM $\gamma_{\boldsymbol{\theta}}$ so that the ratio $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ is roughly flat in the regions where $\mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D)$ has a significant probability mass, so as to approximate the ideally modified target with $\boldsymbol{\theta}$ and $\mathbf{u}$ completely decoupled. This requires the construction of $\gamma_{\boldsymbol{\theta}}$ so that the resulting pushforward density $m(\mathbf{x}|\boldsymbol{\theta})$, as function of $\mathbf{x}$, is a sufficiently accurate approximation of $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$.

3.2. Relation to IS and Pseudo-marginal methods

Observe that the ratio $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ in Equation (4) defines an unbiased IS estimator for the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ based on the IS density $m(\mathbf{x}|\boldsymbol{\theta})$ when $\mathbf{u} \sim N(\mathbf{0}_D, \mathbf{I}_D)$ and the IS simulation sample size is $n = 1$. This observation is important for at least three reasons. First, it implies that the large literature on IS and related methods for hierarchical models (including, e.g., Shephard and Pitt 1997; Richard and Zhang 2007; Rue, Martino, and Chopin 2009; Durbin and Koopman 2012) can be leveraged to choose suitable IS densities $m(\mathbf{x}|\boldsymbol{\theta})$ in order to construct a map $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$. Specific choices are discussed in more detail in Section 4.

Second, it is well known that IS likelihood estimators such as $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ may have infinite variance and thus become unreliable, in particular in high-dimensional applications (Koopman, Shephard, and Creal 2009). This occurs when the tails of $m(\mathbf{x}|\boldsymbol{\theta})$ are thinner than those of the target $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, making $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ unbounded as a function of $\mathbf{u}$. However, under the modified target (4) the likelihood estimator is combined with the thin-tailed standard normal density in $\mathbf{u}$, which counteracts the potential unboundedness of the IS weight in the $\mathbf{u}$-direction. This enhanced robustness with respect to the infinite-variance problem is also evident in the representation (3) of the target. Affine TMs $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$ result in thin-tailed Gaussian IS densities $m(\mathbf{x}|\boldsymbol{\theta})$ and lead to a Jacobian determinant $|\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})|$ which is constant with respect to $\mathbf{u}$. Consequently, in this case, the tail behavior of (3) with respect to $\mathbf{u}$ will be the same as the tail behavior of $p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$ in $\mathbf{x}$. In the simulation experiments discussed further below, it is shown that the proposed method may produce reliable results even when implemented using IS-densities resulting in very large variance of the IS-weights.

Finally, our proposed approach can be seen as a special case of the pseudo-marginal HMC (PM-HMC) method of Alenlöv, Doucet, and Lindsten (2016). PM-HMC relies on joint HMC sampling of MC estimates of the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ and the random variables $(\mathbf{u}^{(i)})$ used to generate those estimates, where the MC likelihood estimates are the average of $n \geq 1$ simulated IS weights $\{\omega_{\boldsymbol{\theta}}(\mathbf{u}^{(i)})\}_{i=1}^{n}$ as given in Equation (4). Alenlöv, Doucet, and Lindsten (2016) show that for an increasing simulation sample size n, the resulting variance reduction in the MC likelihood estimates lead to an increasing decoupling of $\boldsymbol{\theta}$ and $\{\mathbf{u}^{(i)}\}$ under the PM-HMC target, so that PM-HMC moves to HMC sampling on the marginal space of $\boldsymbol{\theta}$. However, since this also leads to an increase in the size of $\{\mathbf{u}^{(i)}\}$, this is at the expense of increased computational cost per evaluation of the PM-HMC target. In their PM-HMC applications to static models with moderate-dimensional latent variables and a low signal-to-noise ratio, Alenlöv, Doucet, and Lindsten (2016) used $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ and find a sufficiently strong decoupling already for the modest size of n. It is well known, however, that in dynamic high-dimensional latent variable models, such brute-force IS estimators using the prior of $\mathbf{x}$ as IS density, which ignores the information about $\mathbf{x}$ in the data likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, typically suffer from a prohibitively large variance for any practical n (Danielsson 1994). Therefore, we propose to use IS densities with higher fidelity than the prior of $\mathbf{x}$ and then reduce the simulation sample size to $n = 1$, in which case the PM-HMC target has the form as given in Equation (3) or Equation (4).

Alenlöv, Doucet, and Lindsten (2016) also proposed a symplectic integrator for approximating the Hamiltonian transition dynamics (2), which is specifically designed for HMC simulation of target distributions of the form Equation (4) with almost complete decoupling. For the ideally modified target with a complete decoupling (which results when $\mathbf{u} \mapsto \omega_{\boldsymbol{\theta}}(\mathbf{u}) \propto 1$), this integrator reduces to a standard leapfrog integrator in the dynamics of $\boldsymbol{\theta}$, whereas the dynamics of the (typically high-dimensional) $\mathbf{u}$ is simulated exactly. Hence, this integrator, which is referred to below as the ADL integrator, appears to be well suited to be combined with our proposed IS-TM-HMC approach. (For further details on the ADL integrator, see the supplementary material, Section A1.)

4. Specific Choices of $m(\mathbf{x}|\theta)$ and $\gamma_\theta(\mathbf{u})$

As alluded to above, the use of $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ to construct TMs can give satisfactory results in cases where the data

$\mathbf{y}$ are rather uninformative about the latent variables $\mathbf{x}$ (see, e.g., Stan Development Team 2019, sec. 2.5). However, as illustrated in Kleppe (2019), such procedures can produce misleading MCMC results if $\mathbf{y}$ provides significant information about $\mathbf{x}$. An even more challenging situation for $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ is when one or more elements of $\boldsymbol{\theta}$ determine how informative $\mathbf{y}$ is with respect to $\mathbf{x}$ (e.g., σ when $y_i|x_i \sim N(x_i, \sigma^2)$), as this may lead to a funnel-shaped target density. On the other hand, as illustrated by Kleppe (2019), rather "crude" TMs that only roughly reflect the location and scale of $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ can lead to dramatic gains in simulation efficiency and cope with funnel-shaped targets. In the following, two strategies for locating TMs are discussed. Both are well known in the context of IS procedures.

4.1. $m(\mathbf{x}|\theta)$ and $\gamma_\theta(\mathbf{u})$ Derived from Laplace Approximations

Assume that $\mathbf{x} \mapsto \log p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is concave for all admissible $\boldsymbol{\theta}$ (as is the case, e.g., when $\mathbf{x}|\boldsymbol{\theta}$ is Gaussian and $\mathbf{x} \mapsto \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ is concave). Then one way to design a TM $\gamma_{\boldsymbol{\theta}}$ for $\mathbf{x}$ that accounts for the location and scale of $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is to construct it from an IS density $m(\mathbf{x}|\boldsymbol{\theta})$ obtained from a local Gaussian Laplace approximation to $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$. (For the use of such IS densities; see, e.g., Shephard and Pitt 1997.) Such a Laplace approximation obtains from a second-order expansion of $\log[p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})]$ around its mode, such that $m(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{h}_{\boldsymbol{\theta}}, \mathbf{G}_{\boldsymbol{\theta}}^{-1})$, where (Rue, Martino, and Chopin 2009)

$$\begin{aligned} \mathbf{h}_{\boldsymbol{\theta}} &= \arg\max_{\mathbf{x}} \log\left[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})\right], \\ \mathbf{G}_{\boldsymbol{\theta}} &= -\nabla_{\mathbf{x}}^2 \log\left[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}}. \end{aligned} \tag{5}$$

The mode $\mathbf{h}_{\boldsymbol{\theta}}$ can be computed iteratively by using Newton's method, which consists of the recursion $\mathbf{h}_{\boldsymbol{\theta}}^{[j]} = \mathbf{h}_{\boldsymbol{\theta}}^{[j-1]} + (\mathbf{G}_{\boldsymbol{\theta}}^{[j-1]})^{-1}\nabla_{\mathbf{x}} \log\left[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{[j-1]}}$, $(j = 1, 2, \ldots)$, initialized by some guesses $\mathbf{h}_{\boldsymbol{\theta}}^{[0]}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{[0]}$ for $\mathbf{h}_{\boldsymbol{\theta}}$ and $\mathbf{G}_{\boldsymbol{\theta}}$. The matrix $\mathbf{G}_{\boldsymbol{\theta}}^{[j]}$ is the negative Hessian of $\log p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ (or some approximation thereof), evaluated at $\mathbf{h}_{\boldsymbol{\theta}}^{[j]}$. Under the assumed concavity, the (exact) Hessian is positive definite, so using it for computing $\mathbf{G}_{\boldsymbol{\theta}}^{[j]}$ ensures stable convergence of the Newton iterations. Notice that the Laplace approximation based on $\mathbf{h}_{\boldsymbol{\theta}}$ and $\mathbf{G}_{\boldsymbol{\theta}}$ is a function of $\boldsymbol{\theta}$ so that the maximal pointwise accuracy of the approximation requires that the Newton iterations be repeated for any new $\boldsymbol{\theta}$ value.

For a given $\boldsymbol{\theta}$ and a fixed number of Newton iterations $J = 0, 1, 2, \ldots$, this Laplace approximation provides a TM of the following form:

$$\gamma_{\boldsymbol{\theta}}(\mathbf{u}) = \mathbf{h}_{\boldsymbol{\theta}}^{[J]} + \left(\mathbf{L}_{\boldsymbol{\theta}}^{[J]}\right)^{-T}\mathbf{u}, \tag{6}$$

where $\mathbf{L}_{\boldsymbol{\theta}}^{[J]}$ is the lower triangular Cholesky factor of $\mathbf{G}_{\boldsymbol{\theta}}^{[J]}$. The resulting Jacobian determinant of $\gamma_{\boldsymbol{\theta}}$, which is required to evaluate the modified HMC target in its representation (3), is simply given by $|\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| = |\mathbf{L}_{\boldsymbol{\theta}}^{[J]}|^{-1}$. For models that imply that the latent variables in $\mathbf{x}$ under $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$ have Markovian properties, the negative Hessian $\mathbf{G}_{\boldsymbol{\theta}}$ is sparse. This may be exploited by using a sparse numerical Cholesky factorization, which leads to a significant reduction in computational costs, especially in high-dimensional applications (see, e.g., Rue, Martino, and Chopin 2009).

As noted above, the Laplace TM (6) must be recomputed for each of the numerical integrator steps during HMC sampling of the target in Equation (3) or Equation (4), as in each of those steps a new $\boldsymbol{\theta}$ value is visited (see Section 2.1). However, rerunning the Newton algorithm for each such step until convergence of $\mathbf{h}_{\boldsymbol{\theta}}^{[J]}$ to a value close to $\mathbf{h}_{\boldsymbol{\theta}}$ can be computationally costly, especially for poorly chosen initial values $\mathbf{h}_{\boldsymbol{\theta}}^{[0]}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{[0]}$.

It is therefore recommended, if this is possible for the model under consideration, to choose for $\mathbf{h}_{\boldsymbol{\theta}}^{[0]}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{[0]}$ values that result from simple closed-form analytical approximations. For this, we use in the first two of our experiments below the standard Bayesian posterior update formulas for the mean and the precision under Gaussian approximations of the likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ and a Gaussian prior $p(\mathbf{x}|\boldsymbol{\theta})$. (Details on these formulas are found in the supplementary material, Section A2, which also provides further details on the implementation of the Laplace TM for all models considered.) If this is not possible, then choosing some fixed initial iterate, say $\mathbf{h}_{\boldsymbol{\theta}}^{[0]} = \mathbf{0}_D$ may be resorted to. Finally, it should be noted that even if a convergence of $\mathbf{h}_{\boldsymbol{\theta}}^{[J]}$ to a value very close $\mathbf{h}_{\boldsymbol{\theta}}$ is generally preferable, the reduction in computational costs achieved if Newton's recursion is iterated only for a rough approximation to $\mathbf{h}_{\boldsymbol{\theta}}$ may compensate for the associated loss in accuracy of $\mathbf{h}_{\boldsymbol{\theta}}^{[J]}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{[J]}$ in reflecting the location and scale of $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$. The optimal choice of J which leads to a balanced tradeoff between the quality of the HMC MCMC samples (determined by the degree of decoupling between $\boldsymbol{\theta}$ and $\mathbf{u}$) and per evaluation computational cost depends on the model under consideration. Such a choice of J remains an open problem that requires further investigation.

It is interesting to note that our proposed IS-TM-HMC approach using Laplace TMs is closely related to the approximate pseudo-marginal MCMC method of Gómez-Rubio and Rue (2018) (see also Margossian et al. 2020). The latter is based on the conventional Laplace approximation of $p(\mathbf{y}|\boldsymbol{\theta})$ (see, e.g., Tierney and Kadane 1986). This approximation is given by $\omega_{\boldsymbol{\theta}}(\mathbf{0}_D)$ (subject to the implicit approximation of the mode and Hessian). Replacing $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ by $\omega_{\boldsymbol{\theta}}(\mathbf{0}_D)$ in our modified HMC target (4) and then integrating analytically with respect to $\mathbf{u}$ results in the approximate MCMC target of Gómez-Rubio and Rue (2018). Therefore, our approach can be seen as an extension of the Gómez-Rubio and Rue (2018) method which consists in correcting the error in their approximate MCMC target.

4.2. $m(\mathbf{x}|\theta)$ and $\gamma_\theta(\mathbf{u})$ Derived from the Efficient Importance Sampler

As an alternative to the Gaussian Laplace approximation to $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$, we consider the efficient importance sampling (EIS) method of Richard and Zhang (2007) for the construction of TMs. While Laplace IS densities are local approximations of $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$, EIS constructs IS densities by least-square (LS) regressions that aim at globally approximating $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ on its full support.

For the application of EIS to the present context, it is assumed that there exists a partition of $\mathbf{x}$ and $\mathbf{y}$ into $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and

$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, so that the conditional likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ and the prior $p(\mathbf{x}|\boldsymbol{\theta})$ can be factorized into low-dimensional densities according to $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{t=1}^{N} p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta})$ and $p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{N} p(\mathbf{x}_t|\mathbf{x}_{(t-1)}, \boldsymbol{\theta})$, where $\mathbf{x}_{(t')} = (\mathbf{x}_1, \ldots, \mathbf{x}_{t'})$ and $p(\mathbf{x}_1|\mathbf{x}_{(0)}, \boldsymbol{\theta}) \equiv p(\mathbf{x}_1|\boldsymbol{\theta})$. Such factorizations can be found for a broad class of Bayesian hierarchial models, including dynamic state-space models (SSMs) for time series. Conformably with the factorization of the prior $p(\mathbf{x}|\boldsymbol{\theta})$, the IS density is decomposed into $m(\mathbf{x}|\mathbf{a}) = \left[\prod_{t=2}^{N} m_t(\mathbf{x}_t|\mathbf{x}_{(t-1)}, \mathbf{a}_t)\right] m_1(\mathbf{x}_1|\mathbf{a}_1)$, with conditional densities m_t such that

$$m_t(\mathbf{x}_t|\mathbf{x}_{(t-1)}, \mathbf{a}_t) = \frac{\tilde{m}_t(\mathbf{x}_{(t)}, \mathbf{a}_t)}{\chi_t(\mathbf{x}_{(t-1)}, \mathbf{a}_t)},$$
$$\chi_t(\mathbf{x}_{(t-1)}, \mathbf{a}_t) = \int \tilde{m}_t(\mathbf{x}_{(t)}, \mathbf{a}_t) d\mathbf{x}_t, \qquad (7)$$

where $\mathcal{M} = \{\tilde{m}_t(\cdot, \mathbf{a}_t), \mathbf{a}_t \in \mathcal{A}_t\}$, is a preselected parametric class of density kernels indexed by auxiliary parameters $\mathbf{a}_t$ and with integrating factors χ_t which are pointwise computable. As required for the proposed TM HMC approach, it is assumed that the IS density $m(\mathbf{x}|\mathbf{a})$ as defined in Equation (7) provides a smooth, invertible mapping $\mathbf{x} = \gamma_{\boldsymbol{\theta},\mathbf{a}}(\mathbf{u})$ with $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_N) \sim N(\mathbf{0}_D, \mathbf{I}_D)$ for any admissible $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_N)$, which is obtained recursively as

$$\mathbf{x}_1 = \gamma_{\boldsymbol{\theta},\mathbf{a}_1}(\mathbf{u}_1), \qquad \mathbf{x}_t = \gamma_{\boldsymbol{\theta},\mathbf{a}_t}(\mathbf{x}_{(t-1)}, \mathbf{u}_t), \qquad t = 2, \ldots, N. \qquad (8)$$

With the assumed factorizations of $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, $p(\mathbf{x}|\boldsymbol{\theta})$ and $m(\mathbf{x}|\mathbf{a})$ the IS weights can be written as

$$\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})}{m(\mathbf{x}|\mathbf{a})} = \chi_1(\mathbf{a}_1) \prod_{t=1}^{N} \times \left[\frac{p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta})p(\mathbf{x}_t|\mathbf{x}_{(t-1)}, \boldsymbol{\theta})\chi_{t+1}(\mathbf{x}_{(t)}, \mathbf{a}_{t+1})}{\tilde{m}_t(\mathbf{x}_{(t)}, \mathbf{a}_t)}\right],$$
$$\chi_{N+1}(\cdot) \equiv 1, \qquad (9)$$

and in order to get a close approximation of $m(\mathbf{x}|\mathbf{a})$ to $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$, EIS aims at selecting a value for $\mathbf{a}$ that sequentially minimizes for each t the variance of the ratio given in brackets of Equation (9) with respect to $m(\mathbf{x}|\mathbf{a})$. For an approximate solution to this minimization problem, EIS solves for the preselected parametric class of kernels $\mathcal{M}$ the following back-recursive sequence of LS problems:

$$(\hat{c}_t, \hat{\mathbf{a}}_t) = \arg\min_{\mathbf{a}_t, c_t} \sum_{i=1}^{r} \times \left\{\log\left[p(\mathbf{y}_t|\mathbf{x}_t^{(i)}, \boldsymbol{\theta})p(\mathbf{x}_t^{(i)}|\mathbf{x}_{(t-1)}^{(i)}, \boldsymbol{\theta})\chi_{t+1}(\mathbf{x}_{(t)}^{(i)}, \hat{\mathbf{a}}_{t+1})\right] - c_t - \log \tilde{m}_t\left(\mathbf{x}_{(t)}^{(i)}, \mathbf{a}_t\right)\right\}^2, \quad t = N, N-1, \ldots, 1, \quad (10)$$

where c_t represents an intercept, and $\{\mathbf{x}^{(i)}\}_{i=1}^{r}$ denote r iid draws simulated from $m(\mathbf{x}|\mathbf{a})$ itself. Thus, the optimal values for the auxiliary parameters $\hat{\mathbf{a}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_N)$ result as a fixed-point solution to the sequence $\{\hat{\mathbf{a}}^{[0]}, \hat{\mathbf{a}}^{[1]}, \ldots\}$ in which $\hat{\mathbf{a}}^{[b]}$ is obtained from Equation (10) under draws from $m(\mathbf{x}|\hat{\mathbf{a}}^{[b-1]})$. In order to ensure convergence to a fixed-point solution all the $\mathbf{x}$ draws simulated for the sequence $\{\hat{\mathbf{a}}^{[b]}\}$ must be generated by using one single set of Gaussian (common) random numbers (CRNs) $\{\mathbf{z}^{(i)}\}_{i=1}^{r}$ with $\mathbf{z}^{(i)} = (\mathbf{z}_1^{(i)}, \ldots, \mathbf{z}_N^{(i)}) \sim$ iid $N(\mathbf{0}_D, \mathbf{I}_D)$ so as to transform them according to Equation (8) into $\mathbf{x}^{(i)} = \gamma_{\boldsymbol{\theta},\hat{\mathbf{a}}^{[b]}}(\mathbf{z}^{(i)})$. Since $\hat{\mathbf{a}}$ is implicitly a function of $\boldsymbol{\theta}$, EIS optimality requires reruns of the LS regressions (10) for any new $\boldsymbol{\theta}$ value. The EIS TM for a given $\boldsymbol{\theta}$ and B EIS fixed-point iterations is then obtained by substituting in Equation (8) the EIS value $\hat{\mathbf{a}}^{[B]}$ for $\mathbf{a}$.

Assume (as is the case in our applications below) that $\mathbf{x}_t \mapsto \log p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta})$ is concave and that $\mathbf{x}$ is Gaussian with a Markovian structure so that $p(\mathbf{x}_t|\mathbf{x}_{(t-1)}, \boldsymbol{\theta}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta})$. Then a natural choice for the $\tilde{m}_t$'s is to use Gaussian density kernels for $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. For this choice the EIS approximation problems (10) take the form of simple low-dimensional linear LS problems. (The corresponding details are provided in the supplementary material, Section A3.1; EIS implementations for applications where $\mathbf{x}$ has priors which are multimodal or exhibit a non-Markovian structure are found in Kleppe and Liesenfeld (2014) and Liesenfeld, Richard, and Vogler (2017). For the selection of corresponding starting values $\hat{\mathbf{a}}^{[0]}$, it is recommended to use values which result from second-order expansions of $\log[p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta})p(\mathbf{x}_t|\mathbf{x}_{(t-1)}, \boldsymbol{\theta})\chi_{t+1}(\mathbf{x}_{(t)}, \hat{\mathbf{a}}_{t+1}^{[0]})]$.

Notice that the EIS values for the parameters of the IS density $\hat{\mathbf{a}}^{[B]}$ are random variables as they depend via the LS regressions (10) on the CRNs $\{\mathbf{z}^{(i)}\}$. This calls for specific rules for the implementation of EIS which ensure that the resulting map $\gamma_{\boldsymbol{\theta},\hat{\mathbf{a}}^{[B]}}$ meets the qualification required for its use as a TM for HMC simulation. First, the CRNs $\{\mathbf{z}^{(i)}\}$ must kept fixed during each HMC update step k with its L numerical integrator steps in which the TM is evaluated for different $\boldsymbol{\theta}$ values. This together with the number of fixed-point iterations B, which is fixed across all $\boldsymbol{\theta}$ values, ensures that $\gamma_{\boldsymbol{\theta},\hat{\mathbf{a}}^{[B]}}$ as a function of $\hat{\mathbf{a}}^{[B]}$ is smooth in $\boldsymbol{\theta}$. The fact that B must be fixed for smoothness of $\gamma_{\boldsymbol{\theta},\hat{\mathbf{a}}^{[B]}}$ is a limitation since it can be expected that the convergence speed of the EIS fixed-point iterations depends on $\boldsymbol{\theta}$. Second, the CRNs $\{\mathbf{z}^{(i)}\}$ need to be included in the Markov kernel defined by the HMC update step, so that it leaves the target posterior $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ invariant and results in a valid HMC. This is achieved by drawing for each HMC update step k a new set of CRNs $\{\mathbf{z}^{(i)}\}$. (For a discussion of the validity, see Section A3.2 of the supplementary material).

With regard to the choice of the EIS tuning parameters r and B, there is a similar tradeoff between the per evaluation computational costs and the degree of decoupling between $\boldsymbol{\theta}$ and $\mathbf{u}$ as for the choice of the Laplace tuning parameter J. In our applications below, for which good starting values $\hat{\mathbf{a}}^{[0]}$ are readily available, the tradeoff suggest only a few EIS iterations B ($B \leq 2$), and r being around two times the number of estimated parameters in Equation (10). However, further investigations are required in order to analyze the optimal choice of (r, B) for any given model, and also to explore the prerequisites that would allow dynamic ($\boldsymbol{\theta}$-dependent) choices of B.

In applications to high-dimensional SSMs, it has been shown that the EIS approach, with the aim of approximating $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ globally, provides IS densities that are more efficient in terms of the variance of the IS weights than those resulting from local Laplace approximations (Bos 2012; Kleppe and Skaug 2012; Koopman, Lucas, and Scharth 2015). This higher fidelity of the EIS IS densities comes at the expense of higher computational costs, and whether it is worthwhile will be examined in our simulation experiments below.

4.3. Implementation and Tuning Parameters

In our simulation experiments below, we consider the following four versions of the proposed IS-TM-HMC approach:

- Stan-Laplace: HMC with the Laplace TM for simulating the target as given in Equation (3) using the leapfrog integrator; This we implemented using Stan.
- ADL-Laplace: HMC with the Laplace TM for simulating the target (3) using the ADL-integrator (as described in Section A1 of the supplementary material).
- ADL-EIS: HMC with the EIS TM for simulating the target (4) using the ADL-integrator.
- Stan-Prior: HMC with a TM, which is defined by $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$, for simulating the target (3) with the leapfrog integrator; This we also implemented in Stan.

Stan-Laplace, ADL-Laplace, and ADL-EIS are the main focus of the article. The brute-force Stan-prior version with a TM based on the prior $p(\mathbf{x}|\boldsymbol{\theta})$ is used as a benchmark for the Laplace and EIS TMs that are designed to take into account the geometry of $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$. The comparison of Stan-Laplace and ADL-Laplace allows us to examine the benefits of moving from the leapfrog to the ADL integrator that is designed for HMC targets with almost complete decoupling. We refrained from an EIS-TM-HMC implementation in Stan, as the implementation of the EIS algorithm in Stan has proven to be impractical.

The tuning parameters for the implementation of the ADL versions (ADL-Laplace, ADL-EIS) are chosen as follows: For the mass matrix in the Hamiltonian (1) with $\mathbf{q} = (\boldsymbol{\theta}^T, \mathbf{u}^T)^T$ and $\mathbf{p} = (\mathbf{p}_{\boldsymbol{\theta}}^T, \mathbf{p}_{\mathbf{u}}^T)^T$, we use

$$\mathbf{M} = \begin{bmatrix} \hat{\mathbf{M}}_{\boldsymbol{\theta}} & \mathbf{0}_{d\times D} \\ \mathbf{0}_{D\times d} & \mathbf{I}_D \end{bmatrix}. \tag{11}$$

Here, the mass matrix specific to $\boldsymbol{\theta}$, $\hat{\mathbf{M}}_{\boldsymbol{\theta}}$, is an approximation to $-\nabla_{\boldsymbol{\theta}}^2 \log\left[\hat{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ (which in turn represents an approximation to the precision matrix of $p(\boldsymbol{\theta}|\mathbf{y})$), where $\hat{p}(\mathbf{y}|\boldsymbol{\theta})$ is a high-precision IS estimate of $p(\mathbf{y}|\boldsymbol{\theta})$ based on EIS, and $\hat{\boldsymbol{\theta}}$ is the simulated maximum posterior probability (MAP) value $\arg\max_{\boldsymbol{\theta}} \log\left[\hat{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right]$. Obtaining $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{M}}_{\boldsymbol{\theta}}$ is very cheap and requires minimal additional coding effort as the gradients of $\log(\hat{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$ are readily available via the automatic differentiation (AD) (Hogan 2014). The columns of the mass matrix $\hat{\mathbf{M}}_{\boldsymbol{\theta}}$ obtain as first order finite difference approximations applied to the AD-based gradient of $\log(\hat{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))$. The mass matrix specific to $\mathbf{u}$ is set equal to the identity so that it matches the precision of the $N(\mathbf{0}_D, \mathbf{I}_D)$ density for $\mathbf{u}$ in Equations (3) and (4). The complete mass matrix (11) is therefore an approximation to the optimal matrix under a Gaussian posterior $p(\boldsymbol{\theta}|\mathbf{y})$ and perfect decoupling of $\boldsymbol{\theta}$ and $\mathbf{u}$. As for the integrator step size ε and the number of integrator steps L, we use L for tuning, while the total integration time εL for an HMC update step is set to approximately $\pi/2$. This choice is informed by the expectation that $\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ in Equation (3) and (4) is nearly Gaussian with a precision matrix (11). Moreover, whenever $\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ is Gaussian with precision (11), the Hamiltonian dynamics according to Equation (2) is periodic with period $t = 2\pi$, and choosing a quarter of such a cycle leads to an AR proposal $\mathbf{q}^*$ in the HMC update step which is independent from the previous HMC draw $\mathbf{q}^{(k)}$ (Neal 2011; Mannseth, Kleppe, and Skaug 2018). Finally, L is tuned such that the acceptance rate for the AR proposals is about 0.9.

For the implementation of all four versions of the IS-TM-HMC, the gradients of the log of the targets (3) and (4) with respect to both $\boldsymbol{\theta}$ and $\mathbf{u}$ are computed using AD. In Stan used for Stan-Laplace and Stan-Prior, this is done automatically and hidden from the user, whereas for ADL-Laplace and ADL-EIS, the Adept C++ AD software library (Hogan 2014) is applied. ADL-Laplace and ADL-EIS are implemented in C++ and interfaced to *R* (R Core Team 2018) using the `Rcpp` (Eddelbuettel and François 2011) package. Stan is used through its *R* interface `rstan` (Stan Development Team 2018), version 2.21.2. The same C++ compiler was used in all implementations. All computations are performed using R version 4.0.2 on a PC with an AMD Ryzen 5 1500X processor running at 3.50 GHz.

5. Simulation Experiments

In this section, we examine the simulation efficiency of our IS-TM-HMC approach for applications to three univariate non-Gaussian/nonlinear SSMs exhibiting different signal-to-noise ratios. In the experiments we analyze the performance of the Stan-Laplace, ADL-Laplace and ADL-EIS versions of the TM-HMC approach and compare them to the brute-force Stan-Prior implementation. Since the SSMs used for the experiments assume that the state innovations $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)$ are standard-normally distributed, Stan-Prior with a TM defined by $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ corresponds to a standard HMC for the $(\boldsymbol{\theta}, \boldsymbol{\eta})$-parameterization. As an additional benchmark method, we consider a modified version of the DRHMC procedure of Kleppe (2019). It uses for HMC simulation of the target in Equation (3) an affine TM of the form

$$\gamma_{\boldsymbol{\theta}}(\mathbf{u}) = \mathbf{h}_F + \mathbf{L}_F^{-T}\mathbf{u}, \quad \mathbf{L}_F\mathbf{L}_F^T = \mathbf{G}_F, \quad \mathbf{G}_F = \mathbf{G}_{\mathbf{x}|\boldsymbol{\theta}} + \mathbf{F}_{\mathbf{y}|\mathbf{x},\boldsymbol{\theta}}, \tag{12}$$

where $\mathbf{G}_{\mathbf{x}|\boldsymbol{\theta}}$ is the precision of the prior $p(\mathbf{x}|\boldsymbol{\theta})$ and $\mathbf{F}_{\mathbf{y}|\mathbf{x},\boldsymbol{\theta}}$ the Fisher information of the data likelihood $p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$ with respect to $\mathbf{x}$ given by $-E[\nabla_{\mathbf{x}}^2 \log p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$. While the original DRHMC uses TMs, which are non-trivial for $\boldsymbol{\theta}$ as well as $\mathbf{x}$, the TM of the modified version considered here is non-trivial only for $\mathbf{x}$. This procedure we have implemented in Stan and is referred to below as Stan-Fisher.

For the purpose of comparing the competing methods, we use the effective sample size (ESS) (Geyer 1992) and the ESS per second of CPU time (ESS/s).

We run the TM-HMC algorithms implemented with the ADL integrator for 10,500 MCMC iterations, where the first 500 are discarded as burn-in. For the algorithms implemented with the leapfrog integrator in Stan, 11,000 iterations with 1000 burn-in steps are used. There the burn-in iterations are also used for automatic tuning of ε (integrator step size) and $\mathbf{M}$ (mass

Table 1. Results for the posterior analysis of the SV model in Equations (13) and (14).

		ADL-EIS		Stan-Prior		ADL-Laplace		Stan-Laplace		Stan-Fisher		PMMH	
		Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean
	CPU time (s)	4023	4031	241	263	137	137	166	180	144	145	639.4	655.7
	$\log(\omega)$ std.		2.2		193		3.0		61				
	$\log(\omega)$ iESS		5.9			1			3.5			1	
γ													
	Post. mean		−0.021		−0.021		−0.021		−0.021		−0.021		−0.021
	Post. std.		0.011		0.011		0.011		0.011		0.011		0.011
	ESS	2597	3104	1862	2451	2546	3185	3286	4481	2624	3219	183	224
	ESS/s	0.6	0.8	6.7	9.5	18.6	23.3	19.8	24	9.8	18.5	0.3	0.3
δ													
	Post. mean		0.98		0.98		0.98		0.98		0.98		0.98
	Post. std.		0.01		0.01		0.01		0.01		0.01		0.01
	ESS	2131	3401	1750	2186	2697	3266	3032	4208	2451	3111	159	195
	ESS/s	0.5	0.8	5.5	8.5	19.7	23.9	18.2	22.4	9.3	17.9	0.2	0.3
v													
	Post. mean		0.15		0.15		0.15		0.15		0.15		0.15
	Post. std.		0.03		0.03		0.03		0.03		0.03		0.03
	ESS	3422	4683	1876	2468	4316	4932	3206	4258	2481	2904	152	213
	ESS/s	0.9	1.2	5.9	9.7	31.5	36	19.1	22.9	9.3	16.7	0.2	0.3

NOTES: ESS is the effective sample size (for 10,000 MCMC draws) and ESS/s is the ESS produced per second of computing time. The figures in the columns "Min" and "Mean" are the values of the minimum and the average across eight independent replications of the algorithms. The tuning parameters are $(B, r, \varepsilon, L) = (2, 6, 0.4, 4)$ for ADL-EIS, $(J, \varepsilon, L) = (2, 0.4, 4)$ for ADL-Laplace, and $J = 0$ for Stan-Laplace. $\log(\omega)$ std. (iESS) is the standard deviation (importance sample effective sample size per 1000 samples) for the log of the IS weights $\omega_{\boldsymbol{\theta}}$ in Equation (4) computed at $(\gamma, \delta, v) = (-0.021, 0.98, 0.14)$.

matrix). For the tuning parameters (L, J) of the ADL-Laplace, (L, B, r) of ADL-EIS and J of Stan-Laplace, we tried different settings. The results reported below are those found for the respective settings that produced the largest ESS/s values. The reported computing times refer to 10,000 MCMC iterations for all implementations. The assumed priors for the parameters of the three example models together with model specific details related to the implementation of the TMs are provided in the supplementary material, Section A4.

5.1. Stochastic Volatility Model

The first example model is the discrete-time stochastic volatility (SV) model for financial returns given by Taylor (1986)

$$y_t = \exp(x_t/2)e_t, \quad e_t \sim \text{iid } N(0,1), \quad t = 1, \dots, N, \qquad (13)$$

$$x_t = \gamma + \delta x_{t-1} + v\eta_t, \quad \eta_t \sim \text{iid } N(0,1), \quad t = 2, \dots, N, \qquad (14)$$

where y_t is the return observed on day t, x_t is the latent log-volatility with initial condition $x_1 \sim N(\gamma/[1-\delta], v^2/[1-\delta^2])$, while e_t and η_t are mutually independent innovations. The data consist of daily log-returns on the U.S. dollar against the U.K. Pound Sterling from October 1, 1981 to June 28, 1985 with $N = 945$.

Under this model, the data density $p(y_t|x_t, \boldsymbol{\theta}) = \mathcal{N}(y_t|0, \exp\{x_t\})$ is fairly uninformative about the states x_t, with a Fisher information (with respect to x_t) which is independent of $\boldsymbol{\theta}$ and given by $-E[\nabla^2_{x_t} \log p(y_t|x_t)] = 1/2$, whereas the states are fairly volatile under typical estimates for $\boldsymbol{\theta}$. The resulting low signal-to-noise ratio together with a shape of the data density which is independent of the parameters implies that the conditional posterior of the state innovation vector $\boldsymbol{\eta}$ given $\boldsymbol{\theta}$ is close to a normal distribution regardless of $\boldsymbol{\theta}$. This leads to a correspondingly well-behaved joint posterior of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, so that a comparatively good performance of the Stan-Prior benchmark sampling on the joint space of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ can be expected. For the TM of Stan-Fisher in (12), we use $\mathbf{h}_F = \mathbf{0}_N$ (as suggested by the results of Kleppe 2019), and the scaling $\mathbf{G}_F$ corresponds to the value of the Bayesian update formula which is used for $\mathbf{G}^{[0]}_{\boldsymbol{\theta}}$ to initialize the Newton search for the Laplace TM.

Table 1 shows the MCMC posterior mean and standard deviation of the parameters for the TM-HMC procedures. They are the sample averages computed from eight independent replications running the algorithms under eight different seeds. Also reported are the corresponding sample average and minimum of the ESS and ESS/s values. For comparison, Table 1 also provides the results for the pseudo-marginal Metropolis-Hastings (PMMH) algorithm implemented in the Bayesian software LibBi (Murray 2015). PMMH is an MH procedure targeting the marginal posterior of the parameters and uses a particle filter for marginalizing the latent variables (Andrieu, Doucet, and Holenstein 2010). The PMMH results are based on 11,000 MH iterations where the first 1000 are discarded (see the supplementary material, Section A4.1 for further details).

It is seen from Table 1 that the five TM-HMC methods produce MCMC estimates for the posterior moments of the parameters that are very close to each other and the ESS values indicate that they all explore the marginal posterior of the parameters well. The ESS values also show that not much is gained by moving from the prior TM to the Fisher, Laplace, or the EIS TM. This finding is in agreement with the results reported in Kleppe (2019) and was to be expected, since the data density is fairly uninformative regarding the states. To assess the fidelity of the IS densities used for constructing the TMs, Table 1 provides the standard deviations for the log of the simulated IS weights $\omega_{\boldsymbol{\theta}}$ (see Equation 4) and the (IS) effective sample size per 1000 draws (iESS) (see, e.g., Doucet and Johansen 2011). They show that ADL-Laplace and ADL-EIS with the ADL-integrator require TMs from IS densities with a much higher fidelity than Stan-Laplace with the leapfrog integrator in order to achieve their ESS/s optimum. The lowest fidelity we observe for the brute-force IS density of Stan-Prior with a standard

Table 2. Results for the posterior analysis of the Gamma model in Equations (15) and (16).

		ADL-EIS		Stan-Prior		ADL-Laplace		Stan-Laplace	
		Min	Mean	Min	Mean	Min	Mean	Min	Mean
	CPU time (s)	13,207	13,226	2112	2167	755	758	699	833
	$\log(\omega)$ std.	2.8		25,816		3.2		14	
	$\log(\omega)$ iESS	5.7		1		16		1	
τ									
	Post. mean		0.13		0.13		0.13		0.13
	Post. std.		0.006		0.006		0.006		0.006
	ESS	8345	9793	1652	1745	10,000	10,000	7002	8613
	ESS/s	0.6	0.7	0.7	0.8	13.1	13.2	9.6	10.5
β									
	Post. mean		2.7		2.8		2.6		2.8
	Post. std.		0.8		0.9		0.8		1
	ESS	3404	4652	880	2057	1379	5170	286	2823
	ESS/s	0.3	0.4	0.4	1	1.8	6.8	0.4	3.3
δ									
	Post. mean		0.98		0.98		0.98		0.98
	Post. std.		0.004		0.004		0.004		0.004
	ESS	4588	5633	1830	2284	3685	6196	1855	6996
	ESS/s	0.3	0.4	0.9	1.1	4.9	8.2	2.6	8.3
ν									
	Post. mean		0.22		0.22		0.22		0.22
	Post. std.		0.01		0.01		0.01		0.01
	ESS	7886	9710	1206	1359	10,000	10,000	5656	7698
	ESS/s	0.6	0.7	0.6	0.6	13.1	13.2	8.1	9.2

NOTES: ESS is the effective sample size (for 10,000 MCMC draws) and ESS/s is the ESS produced per second of computing time. The figures in the columns "Min" and "Mean" are the values of the minimum and the average across eight independent replications of the algorithms. The tuning parameters are $(B, r, \varepsilon, L) = (2, 5, 0.64, 3)$ for ADL-EIS, $(J, \varepsilon, L) = (1, 0.64, 3)$ for ADL-Laplace, and $J = 0$ for Stan-Laplace. $\log(\omega)$ std. (iESS) is the standard deviation (importance sample effective sample size per 1000 samples) for the log of the IS weights $\omega_{\boldsymbol{\theta}}$ in Equation (4) computed at $(\tau, \beta, \delta, \nu) = (0.13, 2.8, 0.98, 0.22)$.

deviation of the log IS weights as large as 193 and an iESS value as low as one (indicating that to machine precision, a single importance weight will account for the complete IS estimate based on 1000 samples). The good performance of Stan-prior and Stan-Laplace, achieved even at low-fidelity IS densities, is in sharp contrast to the performance that can be expected in standard IS application to the marginalization of latent variables. There such IS densities with standard deviations of the log IS weight of the order of say 10 and larger are useless for accurately approximating, for example, a marginal likelihood. With regard to ESS/s, there is no method that is consistently the best, but it turns out that the computational overhead for positioning the EIS TM is clearly not worthwhile for this model compared to the cheaper construction of the Laplace and Fisher TMs. Finally, we find that PMMH is clearly outperformed by the other methods in terms of ESS. Since PMMH was run on a different (though similar) computer than the other methods, its reported ESS/s values are only approximately comparable to the other ESS/s figures. Still, the substantial differences in the ESS/s values suggest that PMMH cannot compete with the TM-HMC methods based on the Laplace and prior TM in terms of simulation efficiency.

5.2. Gamma Model for Realized Volatilities

The second example model is a dynamic SSM for the realized variance of asset returns (see, e.g., Golosnoy, Gribisch, and Liesenfeld 2012, and references therein). It has the form

$$y_t = \beta \exp(x_t) e_t, \quad e_t \sim \text{iid } G(1/\tau, \tau), \quad t = 1, \ldots, N, \quad (15)$$
$$x_t = \delta x_{t-1} + \nu \eta_t, \quad \eta_t \sim \text{iid } N(0, 1), \quad t = 2, \ldots, N, \quad (16)$$

where y_t is the daily realized variance measuring the latent integrated variance $\beta \exp(x_t)$, and $G(1/\tau, \tau)$ denotes a Gamma-distribution for e_t normalized such that $E(e_t) = 1$ and $\text{var}(e_t) = \tau$. The innovations e_t and η_t are independent and the initial condition for the log-variance is $x_1 \sim N(0, \nu^2/[1 - \delta^2])$. The data are taken from Golosnoy, Gribisch, and Liesenfeld (2012) and consists of daily realized variances for the American Express stock from January 1, 2000 to December 31, 2009 with $N = 2,514$.

In contrast to the SV model, the Gamma model has both a considerably higher signal-to-noise ratio and a shape of the data density $p(y_t|x_t, \boldsymbol{\theta})$ which depends on the parameters. In particular, the Fisher information of its data density is $1/\tau$ with an estimate of $\tau \simeq 0.13$ (see Table 2), while the estimated volatility of the states is roughly as large as under the SV model. Hence, it can be expected that the conditional posterior of the innovations $\boldsymbol{\eta}$ given $\boldsymbol{\theta}$ deviates distinctly from a Gaussian distribution and exhibits nonlinear dependence on $\boldsymbol{\theta}$, which makes the Gamma model a more challenging scenario for the Stan-Prior benchmark than the SV model.

As for the SV model, we use the Bayesian update formulas for $\mathbf{h}_{\boldsymbol{\theta}}^{[0]}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{[0]}$ in the Laplace-TM implementations, and again the resulting $\mathbf{G}_{\boldsymbol{\theta}}^{[0]}$ is equal to the scaling $\mathbf{G}_F$ in the Fisher TM. Since the Fisher TM with $\mathbf{h}_F = \mathbf{0}_N$ leads to poor results, we set $\mathbf{h}_F = \mathbf{h}_{\boldsymbol{\theta}}^{[0]}$ (see also Kleppe 2019, eq. 20). As a result, Stan-Fisher corresponds to Stan-Laplace with $J = 0$. This number of Newton iterations also turned out to be optimal for Stan-Laplace in terms of ESS/s.

The results of the experiment for the Gamma model are summarized in Table 2. They reveal a significant improvement in the ESS when replacing the brute-force prior TM by the Laplace or EIS TM. This improvement is mainly due to the substantial amount of information in the data density about the states that the EIS and Laplace TM take into account in modifying the target. In terms of ESS/s, ADL-Laplace, and Stan-Laplace are the best performing methods, and again it is not beneficial for the TM design to move from the Laplace approximation to the presumably more accurate but computationally more expensive EIS approximation. Finally, we observe that, as in the SV model, Stan-Laplace gives reliable MCMC results even with a TM from an IS density with a fairly large variance in the IS weight, while ADL-Laplace requires an IS density with higher fidelity.

5.3. Constant Elasticity of Variance Diffusion Model

The third example model is a time-discretized version of the constant elasticity of variance (CEV) diffusion model for interest rates (Chan et al. 1992), extended by a measurement error to account for microstructure noise (Aït-Sahalia 1999). Under this model the interest rate y_t observed at day t with a corresponding latent state $x_t > 0$, is described as

$$y_t = x_t + \sigma_y e_t, \quad e_t \sim \text{ iid } N(0,1), \quad t = 1,\dots,N, \tag{17}$$

$$x_t = x_{t-1} + \Delta(\alpha - \beta x_{t-1}) + \sigma_x x_{t-1}^{\gamma}\sqrt{\Delta}\eta_t,$$
$$\eta_t \sim \text{ iid } N(0,1), \quad t = 2,\dots,N, \tag{18}$$

where e_t and η_t are mutually independent and $\Delta = 1/252$. The initial condition assumed for the state is $x_1 \sim N(y_1, 0.01^2)$. The data consist of $N = 3082$ daily 7-day Eurodollar deposit spot rates from January 2, 1983 to February 25, 1995.

The estimated standard deviation of the measurement error σ_y is very small with an estimate of 0.0005 (see Table 3) so that the data density $x_t \mapsto p(y_t|x_t, \boldsymbol{\theta})$ is strongly peaked at $x_t = y_t$ and by far more informative about x_t than in the SV and Gamma model with a Fisher information given by $1/\sigma_y^2$. Also, the volatility of the states is not constant and depends, unlike in the previous models, nonlinearly on the level of the states. This leads to a posterior for $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ which strongly deviates from a Gaussian distribution. As a result, Stan-Prior fails to produce meaningful MCMC results so that we refrain from reporting them. Moreover, since the prior of $\mathbf{x}$ is nonlinear and its precision matrix does not have closed-form, the use of the Fisher TM in (12) is not feasible.

Table 3 reports the results for ADL-EIS, ADL-Laplace, and Stan-Laplace. It is seen that they all lead to a reliable exploration of the marginal posterior of the parameters with fairly large values of the ESS. For this, ADL-Laplace needs substantially less computing time than ADL-EIS and Stan-Laplace as it is indicated by the ESS/s values. The reason for the observed large differences in the ESS/s values between ADL- and Stan-Laplace is that the leapfrog integrator for Stan-Laplace requires considerably more steps than the ADL-integrator used for ADL-Laplace. Finally, we find that not only ADL-EIS and ADL-Laplace use quite high-fidelity IS densities for the TM but also Stan-Laplace, which is related to that $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is very close to Gaussian distribution.

The CEV model with the same data as here is also considered by Kleppe (2018, sec. 5), who compare the modified Cholesky Riemann manifold HMC algorithm and a Gibbs sampling procedure. Both methods were implemented in C ++, and therefore the order of magnitudes of the ESS/s values reported there are comparable to those given in Table 3. The ESS/s comparison

Table 3. Results for the posterior analysis of the CEV model in Equations (17) and (18).

		ADL-EIS		ADL-Laplace		Stan-Laplace	
		Min	Mean	Min	Mean	Min	Mean
	CPU time (s)	36,640	45,201	901	902	9413	9458
	log(ω) std.		0.3		0.5		3.2
	log(ω) iESS		930		812		13
α							
	Post. mean		0.01		0.01		0.01
	Post. std.		0.01		0.01		0.01
	ESS	8974	9658	9305	9748	10,000	10,000
	ESS/s	0.2	0.2	10.3	10.8	1.1	1.1
β							
	Post. mean		0.17		0.17		0.17
	Post. std.		0.17		0.17		0.17
	ESS	9471	9872	9330	9801	10,000	10,000
	ESS/s	0.2	0.2	10.3	10.9	1.1	1.1
γ							
	Post. mean		1.18		1.18		1.18
	Post. std.		0.06		0.06		0.06
	ESS	8778	9408	9614	9930	7009	9134
	ESS/s	0.2	0.2	10.7	11	0.7	1
σ_x							
	Post. mean		0.41		0.41		0.41
	Post. std.		0.06		0.06		0.06
	ESS	8356	9334	9263	9815	7235	9295
	ESS/s	0.2	0.2	10.3	10.9	0.8	1
σ_y							
	Post. mean		0.0005		0.0005		0.0005
	Post. std.		0.00002		0.00002		0.00002
	ESS	9211	9508	9316	9711	10,000	10,000
	ESS/s	0.2	0.2	10.3	10.8	1.1	1.1

NOTES: ESS is the effective sample size (for 10,000 MCMC draws) and ESS/s is the ESS produced per second of computing time. The figures in the columns "Min" and "Mean" are the values of the minimum and the average across eight independent replications of the algorithms. The tuning parameters are $(B, r, \varepsilon, L) = (1, 7, 0.57, 3)$ for ADL-EIS, $(J, \varepsilon, L) = (2, 0.57, 3)$ for ADL-Laplace, and $J = 1$ for Stan-Laplace. log(ω) std. (iESS) is the standard deviation (importance sample effective sample size per 1000 samples) for the log of the IS weights $\omega_{\boldsymbol{\theta}}$ in Equation (4) computed at $(\alpha, \beta, \gamma, \sigma_x, \sigma_y) = (0.01, 0.17, 1.18, 0.41, 0.0005)$.

shows that for the most critical parameters with regard to the ESS (γ and σ_x), ADL-Laplace is about two orders of magnitude faster than the Riemann manifold HMC procedure and about three orders of magnitude faster than the Gibbs sampler.

5.4. Summary of Results

For dynamic SSMs with a signal-to-noise ratio greater than that in the SV model, our proposed IS-TM-HMC approach significantly accelerates the exploration of the marginal posterior of the parameters compared to the benchmarks (or enables a reliable exploration as in the case of the CEV model). Due to its high computational cost, the EIS TM turned out to be not competitive compared to the Laplace TM. For the Laplace-TM approach, when it is implemented in Stan with the leapfrog integrator, a rather low fidelity IS density (as can be obtained from relatively few Newton-search iterations) is optimal in an ESS-per-time-unit perspective. This is partly due to the very flexible tuning of Stan. When implemented with the ADL integrator, which relies on the preselected mass matrix (11), the Laplace-TM procedure requires more accurate IS densities. Overall, and in line with Kleppe (2019), this shows that fairly rough representations of the location and scale of $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ are sufficient when IS-TM-HMC is combined with a flexible tuning of the HMC parameters. In addition, this result illustrates the second point discussed in Section 3.2: Due to the thin-tailed Gaussian distribution entering representation (4) of the modified target, the rule of thumb for standard IS applications that the IS density should provide high-fidelity approximations to $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ is less relevant for our IS-TM-HMC approach.

With respect to the choice of the integrator, it turned out that the leapfrog integrator in Stan and the ADL integrator produce similarly large ESSs for HMC based on the Laplace TM. For this, the ADL integrator requires significantly fewer integrator steps than its leapfrog counterpart. For example, the reported (automatically tuned) Stan-Laplace results for the CEV model required on average about 60 leapfrog steps whereas the corresponding (manually tuned) number of steps for ADL-Laplace was 3.

6. Simulation Experiment for a High-Dimensional Application

6.1. Model

To illustrate the IS-TM-HMC approach in a high-dimensional application, we consider the dynamic inverse-Wishart model for the realized covariance matrix of a set of G asset returns as proposed in (Grothe, Kleppe, and Liesenfeld 2019). It assumes for the $G \times G$ realized covariance matrix $\mathbf{Y}_t$ observed in period t a conditional inverted Wishart distribution with density

$$p(\mathbf{Y}_t|\Sigma_t,\nu) \propto |\mathbf{Y}_t|^{-(\nu+G+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma_t\mathbf{Y}_t^{-1}\right)\right), \quad t = 1,\ldots,N, \tag{19}$$

where Σ_t is a latent time-varying positive-definite scale matrix and $\nu > G+1$ the degrees of freedom. The scale matrix is taken to depend on a Gaussian autoregressive state vector $(x_{1,t},\ldots,x_{G,t})$ in the form

$$\Sigma_t = \mathbf{H}\mathbf{D}_t\mathbf{H}^T, \qquad \mathbf{D}_t = \mathrm{diag}(\exp(x_{1,t}),\ldots,\exp(x_{G,t})), \tag{20}$$

$$x_{g,t} = \mu_g + \delta_g(x_{g,t-1}-\mu_g) + \sigma_g\eta_{g,t}, \qquad \eta_{g,t} \sim \text{ iid } N(0,1), \quad g = 1,\ldots,G, \tag{21}$$

with $x_{g,1} \sim N(\mu_g, \sigma_g^2/(1-\delta_g^2))$. The matrix $\mathbf{H}$ is a lower-triangular matrix with unit diagonal elements and unrestricted parameters $h_{g,\ell}$ ($g > \ell, 1 \leq \ell < G$) below the main diagonal. In total, the model contains $1+3G+G(G-1)/2$ parameters given by $\boldsymbol{\theta} = (\nu,\mu_1,\delta_1,\sigma_1,\ldots,\mu_G,\delta_G,\sigma_G,h_{2,1},\ldots,h_{G,G-1})$.

A fortunate property of this inverse-Wishart SSM is that the G individual univariate state processes $\mathbf{x}_g = (x_{g,1},\ldots,x_{g,N})$, $g = 1,\ldots,G$, are mutually independent under their joint conditional posterior given $\boldsymbol{\theta}$, so that this posterior factorizes into $p(\mathbf{x}|\mathbf{Y},\boldsymbol{\theta}) = \prod_{g=1}^G p(\mathbf{x}_g|\mathbf{Y},\boldsymbol{\theta})$, where $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_G)$ and $\mathbf{Y} = (\mathbf{Y}_1,\ldots,\mathbf{Y}_N)$. The individual conditional posteriors for the G state processes are given by

$$p(\mathbf{x}_g|\mathbf{Y},\boldsymbol{\theta}) \propto p(\mathbf{x}_g|\boldsymbol{\theta})\prod_{t=1}^N \exp\left(\frac{\nu}{2}x_{g,t} - \frac{\tilde{y}_{g,t}}{2}\exp(x_{g,t})\right), \quad \tilde{y}_{g,t} = \mathbf{h}_g^T\mathbf{Y}_t^{-1}\mathbf{h}_g, \quad g = 1,\ldots,G, \tag{22}$$

where $\mathbf{h}_g$ denotes the gth column of the matrix $\mathbf{H}$ and $p(\mathbf{x}_g|\boldsymbol{\theta})$ is the prior of $\mathbf{x}_g$ defined by the Gaussian autoregressive process in Equation (21). This factorization implies that we can construct a TM for $\mathbf{x}$ from an IS density, which is decomposed conformably with the conditional posterior $p(\mathbf{x}|\mathbf{Y},\boldsymbol{\theta})$ into $m(\mathbf{x}|\mathbf{Y},\boldsymbol{\theta}) = \prod_{g=1}^G m_g(\mathbf{x}_g|\boldsymbol{\theta})$, where each m_g is constructed as an approximation to Equation (22). This yields a TM $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$ for $\mathbf{x}$ which is split into G independent maps, one for each state process, say $\mathbf{x}_g = \gamma_{\boldsymbol{\theta},g}(\mathbf{u}_g)$ with $\mathbf{u}_g = (u_{g,1},\ldots,u_{g,N}) \sim N(\mathbf{0}_N,\mathbf{I}_N)$, $g = 1,\ldots,G$. Then the complete TM is $\gamma_{\boldsymbol{\theta}}(\mathbf{u}) = [\gamma_{\boldsymbol{\theta},1}(\mathbf{u}_1)^T,\ldots,\gamma_{\boldsymbol{\theta},G}(\mathbf{u}_G)^T]^T$ and the Jacobian determinant in the modified target (3) obtains as $|\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| = \prod_{g=1}^G |\nabla_{\mathbf{u}_g}\gamma_{\boldsymbol{\theta},g}(\mathbf{u}_g)|$. Obviously, this utilization of the conditional independence of the state processes simplifies the implementation of the IS-TM-HMC approach in this high-dimensional application. However, it should be pointed out that this approach per-se can also be used without such independence.

The Fisher information of the density $p(\tilde{y}_{g,t}|x_{g,t},\boldsymbol{\theta})$ for the standardized realized covariance observations $\tilde{y}_{g,t}$ defined in (22) is $\nu/2$. For the data used in this experiment, the estimate for ν is 33.61 (see supplementary material, Table A1), so the signal-to-noise ratio is similar to that of the fitted Gamma model in Section 5.2. Since the results of ADL- and Stan-Laplace turned out to be similar for the Gamma model, we only consider Stan-Laplace for the inverse-Wishart model, which can be implemented with a few dozen lines of Stan code and fully automated tuning. ADL-EIS was found not to be competitive and is not considered here.

6.2. Results

The data consists of $N = 2,514$ daily realized covariance matrices for $G = 5$ stocks (American Express, Citigroup, General Electric, Home Depot, IBM) spanning January 1, 2000 to

Table 4. ESS for the parameters and the first-period states $x_{g,1}$ and $u_{g,1}$ of the inverse Wishart model in Equations (19)–(21).

		Stan-Prior	Stan-Laplace	Stan-Laplace	Stan-Laplace
			$J = 0$	$J = 1$	$J = 2$
CPU time (s)		142,506	20,067	25,940	31,180
ν	ESS	3750	10,000	10,000	10,000
$h_{g,\ell}$	ESS (min, max)	(8316 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)
μ_g	ESS (min, max)	(2600 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)
δ_g	ESS (min, max)	(2103 , 5205)	(10,000 , 10,000)	(9291 , 10,000)	(10,000 , 10,000)
σ_g	ESS (min, max)	(2372 , 3476)	(10,000 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)
$x_{g,1}$	ESS (min, max)	(9887 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)
$u_{g,1}$	ESS (min, max)	(3059 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)	(10,000 , 10,000)

NOTES: Parameters and states are grouped (with $g = 1, \ldots, 5$ and $\ell < g$, $1 < \ell < 5$), and reported ESS figures are the minimum and maximum in each group. All figures are averages across 8 independent replications of the algorithms. Under the Prior TM, $u_{g,1}$ is identical to $\eta_{g,1}$ in Equation (21).

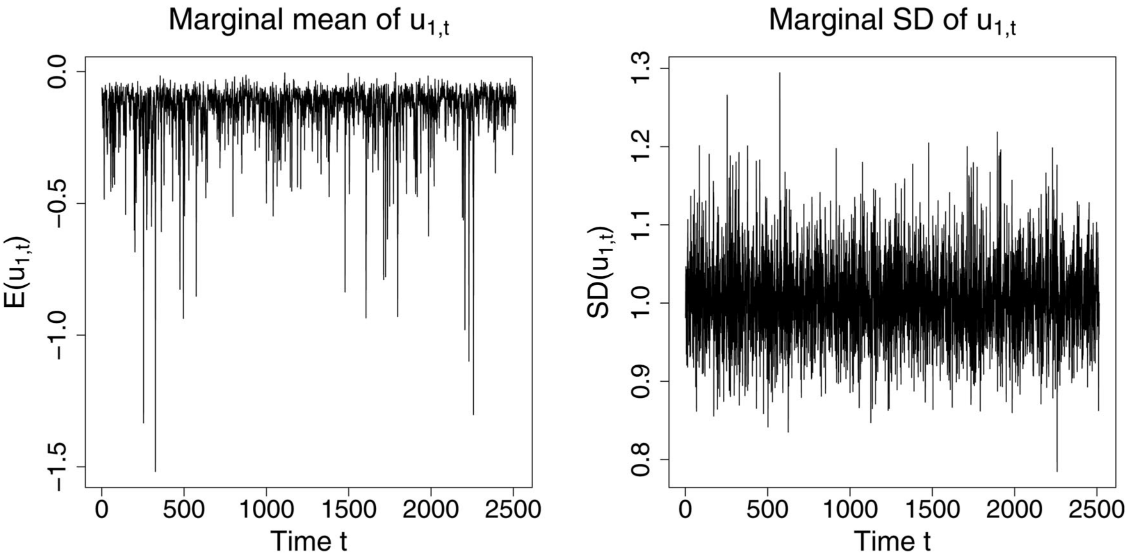

Figure 1. MCMC posterior mean and standard deviation of $\mathbf{u}_1$ for the inverse Wishart model in Equations (19)–(21) under Laplace TM with $J = 0$. The results are for a single representative simulation replication with 1000 MCMC iterations. Corresponding plots for the remaining $\mathbf{u}_g$s may be found in the supplementary material.

December 31, 2009 (Golosnoy, Gribisch, and Liesenfeld 2012). The ESS values for the parameters and the first-period states $\{x_{g,1}\}_{g=1}^5$ and $\{u_{g,1}\}_{g=1}^5$ are found in Table 4. They are reported for Stan-Prior and Stan-Laplace with $J = 0, 1,$ and 2 iterations used for the Newton search of the Laplace TM. This search has been initialized by using the Bayesian update formulas for $\mathbf{h}_{\boldsymbol{\theta},g}^{[0]}$ and $\mathbf{G}_{\boldsymbol{\theta},g}^{[0]}$ (see the supplementary material, Section A4.4, which also provides the MCMC posterior estimates of the parameters). The results in Table 4 show that Stan-Laplace outperforms the Stan-Prior benchmark, both with respect to CPU time and ESS. In fact, the ESS achieved by Stan-Laplace with $J = 0$ per CPU time unit for the parameters ν and $\{\sigma_g\}$, which are the most critical in terms of ESS, is at least one order of magnitude greater than for Stan-Prior. We also see that the additional computational cost of increasing J to one or two to get TMs that more accurately reflect the location and scale of $p(\mathbf{x}|\mathbf{Y}, \boldsymbol{\theta})$ is not worth it. This again corroborates our previous results in Section 5, that for an efficient exploration of the target, it is sufficient if the TM only roughly reflects the geometry of $p(\mathbf{x}|\mathbf{Y}, \boldsymbol{\theta})$.

Figure 1 plots the time series of the MCMC posterior means and standard deviations of each $u_{1,t}$, for Stan-Laplace with $J = 0$. The results for $g = 2, \ldots, G$ are qualitatively similar and are presented in Figure A1 (supplementary material). The posterior means and standard deviations should be zero and one, respectively, if $\boldsymbol{\theta}$ and $\mathbf{u}$ are completely decoupled under the modified target (4). It is seen that the MCMC standard deviations fluctuate very closely around the reference value of one, which indicates that any funnel-shaped form of the original target distribution has been removed by using the Laplace TM to move to the modified target. However, the MCMC means are systematically lower than the reference value. This is due to the fact that, on the one hand, the location parameters $\mathbf{h}_{\boldsymbol{\theta},g}^{[0]}$ of the Laplace TM with $J = 0$ (as they result from the Bayesian update formula) only provide rough approximations to the actual positions of the conditional posteriors $p(\mathbf{x}_g|\mathbf{Y}, \boldsymbol{\theta})$ in Equation (22) and, on the other hand, that these posteriors are non-Gaussian, so that TMs from Gaussian IS density approximations cannot completely decouple $\boldsymbol{\theta}$ and $\mathbf{u}$. If J is increased to $J = 2$ and $J = 10$, the posterior means of $u_{g,t}$ come closer to the reference value of zero but systematically remain below because of the non-Gaussian form of the conditional posteriors (22) (see supplementary material, Figures A2 and A3).

The inverse-Wishart model in Equations (19)–(21) with the data as here is also considered by Grothe, Kleppe, and Liesenfeld (2019). They use a Gibbs sampling procedure that generates MCMC draws from the conditional posterior of the states $p(\mathbf{x}|\mathbf{Y}, \boldsymbol{\theta})$, which are as good as iid. The comparison of the ESS values in Grothe, Kleppe, and Liesenfeld (2019) with those in

Table 4 reveals considerable gains by Stan-Laplace compared to this Gibbs method in exploring the posterior of the parameters, especially for the most critical parameters (ν and $\{\sigma_g\}$). Although the computing times are not directly comparable due to the different programming environments used, a corresponding estimate shows that Stan-Laplace is also about an order of magnitude faster than the Gibbs approach. Equally, important for this comparison is to weigh the minimal coding effort required in Stan to fit a model with 26 parameters and 12,570 latent variables in a few minutes of CPU time against the typically time-consuming and error-prone development of a model-tailored Gibbs algorithm.

7. Discussion

In this article, we have proposed a TM-HMC approach based on IS for Bayesian hierarchical models. This approach uses off-the-shelf IS strategies for high-dimensional latent variables to modify the target distribution so that it can be easily sampled using (fixed metric) HMC. We have illustrated that the proposed approach can significantly accelerate the exploration of the posterior distribution for models with high-dimensional latent variables relative to relevant benchmarks, while being easily implemented using, for example, the software Stan. We considered two high-dimensional IS algorithms (Laplace and EIS), and used them to analyze in simulation experiments the optimal tradeoff for HMC simulation efficiency between the degree of fidelity the IS density exhibits and computational costs. The main insight that was gained from these experiments is that only rather crude IS densities are required when these are combined with the HMC algorithm implemented in Stan. This finding contradicts the general experience with IS documented in the literature that reliable MC estimates used to marginalize high-dimensional latent variables typically require very accurate IS densities.

There are several avenues for future research related to our approach. One would be to generalize the decomposition of a high-dimensional TM into several low-dimensional ones (as we have used it to exploit the independence structure of the inverse-Wishart model in Section 6) to applications with conditional dependence structures (Spantini, Bigoni, and Marzouk 2018). In this way, more complex models, for example, with multiple layers of latent variables, may also be treated by a composite of low-dimensional TMs that are constructed from IS densities. Based on our results regarding the required accuracy of the IS densities, another direction for future research is to consider other methods of approximating the conditional posterior of the latent states $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$. These may include computationally fast variational methods, and in cases where $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ is multimodal, fitting Laplace-type TMs to tempered versions of $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$. Another strategy worth exploring is to fit a normalizing flow model to initial samples generated from $p(\mathbf{x},\mathbf{u},\boldsymbol{\theta}|\mathbf{y})$ by our method (see, e.g., Hoffman et al. 2019). Then this fitted model can be used for the TM construction instead of recalculating the map for each new value of $\boldsymbol{\theta}$.

Finally, we plan to develop software in which the proposed methodology is implemented in a user-friendly manner for a large class of models. In particular, such software should include sparse Cholesky algorithms for more general sparsity structures so that Laplace TMs for multivariate latent state dynamic models and spatial models can also be considered. Further research on how to leverage the ADL integrator in a more automatically tuned manner, similarly to Stan, is also planned.

Acknowledgments

The authors are indebted to the Editor, Tyler McCormick, an anonymous Associate Editor and two anonymous Reviewers, Yushu Li, Christopher Nemeth, Hans Julius Skaug for comments and discussions.

Supplementary Materials

Supplementary Material: Contains detailed formulas for ADL integrator, implementation and model details, and an argument for the correctness of EIS-based methods. (TMHMC_supplementary.pdf).

ORCID

Tore Selland Kleppe http://orcid.org/0000-0001-8469-908X

References

Aït-Sahalia, Y. (1999), "Transition Densities for Interest Rate and Other Nonlinear Diffusions," *The Journal of Finance*, 54, 1361–1395. [10]

Alenlöv, J., Doucet, A., and Lindsten, F. (2016), "Pseudo-Marginal Hamiltonian Monte Carlo," arXiv: 1607.02516. [2,4]

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 72, 269–342. [1,8]

Bos, C. S. (2012), "Relating Stochastic Volatility Estimation Methods," in *Handbook of Volatility Models and Their Applications* (Chapter 6), eds. L. Bauwens, C. Hafner, and S. Laurent. Hoboken, NJ: Wiley, pp. 147–174. [6]

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992), "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate," *The Journal of Finance*, 47, 1209–1227. [10]

Danielsson, J. (1994), "Stochastic Volatility in Asset Prices: Estimation With Simulated Maximum Likelihood," *Journal of Econometrics*, 64, 375–400. [2,4]

Doucet, A., and Johansen, A. (2011), "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later," in *The Oxford Handbook of Nonlinear Filtering* (Chapter 25), eds. D. Crisan and B. Rozovskii, Oxford University Press, pp. 656–704. [8]

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), "Hybrid Monte Carlo," *Physics Letters B*, 195, 216–222. [2]

Durbin, J., and Koopman, S. J. (2012), *Time Series Analysis by State Space Methods* (2nd ed.), Number 38 in Oxford Statistical Science. Oxford: Oxford University Press. [4]

Eddelbuettel, D., and François, R. (2011), "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40, 1–18. [7]

Flury, T., and Shephard, N. (2011), "Bayesian Inference Based Only on Simulated Likelihood: Particle Filter Analysis of Dynamic Economic Models," *Econometric Theory*, 27, 933–956. [1]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. (2014), *Bayesian Data Analysis* (3rd ed.), London: CRC Press. [3]

Geyer, C. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 473–483. [7]

Girolami, M., and Calderhead, B. (2011), "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 73, 123–214. [1]

Golosnoy, V., Gribisch, B., and Liesenfeld, R. (2012), "The Conditional Autoregressive Wishart Model for Multivariate Stock Market Volatility," *Journal of Econometrics*, 167, 211–223. [9,12]

Gómez-Rubio, V., and Rue, H. (2018), "Markov Chain Monte Carlo With the Integrated Nested Laplace Approximation," *Statistics and Computing*, 28, 1033–1051. [5]

Grothe, O., Kleppe, T. S., and Liesenfeld, R. (2019), "The Gibbs Sampler With Particle Efficient Importance Sampling for State-Space Models," *Econometric Reviews*, 38, 1152–1175. [11,12]

Hoffman, M., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. (2019), "Neutra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport," arXiv: 1903.03704. [1,3,13]

Hogan, R. J. (2014), "Fast Reverse-Mode Automatic Differentiation Using Expression Templates in C++," *ACM Transactions on Mathematical Software (TOMS),* 40, Article no. 26. [7]

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994), "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business & Economic Statistics,* 12, 371–89. [1]

Kleppe, T. S. (2018), "Modified Cholesky Riemann Manifold Hamiltonian Monte Carlo: Exploiting Sparsity for Fast Sampling of High-Dimensional Targets," *Statistics and Computing,* 28, 795–817. [1,10]

Kleppe, T. S. (2019), "Dynamically Rescaled Hamiltonian Monte Carlo for Bayesian Hierarchical Models," *Journal of Computational and Graphical Statistics,* 28, 493–507. [1,3,5,7,8,9,11]

Kleppe, T. S., and Liesenfeld, P. E. (2014), "Efficient Importance Sampling in Mixture Frameworks," *Computational Statistics & Data Analysis* 76, 449 – 463. [6]

Kleppe, T. S., and Skaug, H. J. (2012), "Fitting General Stochastic Volatility Models Using Laplace Accelerated Sequential Importance Sampling," *Computational Statistics & Data Analysis*, 56, 3105–3119. [6]

Koopman, S. J., Lucas, A., and Scharth, M. (2015), "Numerically Accelerated Importance Sampling for Nonlinear Non-Gaussian State-Space Models," *Journal of Business & Economic Statistics*, 33, 114–127. [6]

Koopman, S. J., Shephard, N., and Creal, D. (2009), "Testing the Assumptions Behind Importance Sampling," *Journal of Econometrics*, 149, 2 – 11. [2,4]

Leimkuhler, B., and Reich, S. (2004), *Simulating Hamiltonian Dynamics*, Cambridge: Cambridge University Press. [3]

Liesenfeld, R., Richard, J.-F., and Vogler, J. (2017), "Likelihood-Based Inference and Prediction in Spatio-Temporal Panel Count Models for Urban Crimes," *Journal of Applied Econometrics*, 32, 600–620. [6]

Mannseth, J., Kleppe, T. S., and Skaug, H. J., (2018), "On the Application of Improved Symplectic Integrators in Hamiltonian Monte Carlo," *Communications in Statistics–Simulation and Computation*, 47, 500–509. [7]

Margossian, C. C., Vehtari, A., Simpson, D., and Agrawal, R. (2020), "Hamiltonian Monte Carlo using an Adjoint-Differentiated Laplace Approximation," arXiv: 2004.12550. [5]

Murray, L. (2015), "Bayesian State-Space Modelling on High-Performance Hardware Using Libbi," *Journal of Statistical Software, Articles*, 67, 1–36. [8]

Neal, R. M. (2011), "MCMC Using Hamiltonian Dynamics," *Handbook of Markov Chain Monte Carlo*, 2, 113–162. [2,3,7]

Parno, M., and Marzouk, Y. (2018), "Transport Map Accelerated Markov Chain Monte Carlo," *SIAM/ASA Journal on Uncertainty Quantification*, 6, 645–682. [1,3]

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012), "On Some Properties of Markov Chain Monte Carlo Simulation Methods Based on the Particle Filter," *Journal of Econometrics*, 171, 134 – 151. [1]

R Core Team. (2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [7]

Richard, J.-F., and Zhang, W. (2007), "Efficient High-dimensional Importance Sampling," *Journal of Econometrics*,141, 1385–1411. [4,5]

Robert, C., and Casella, G. (2004), *Monte Carlo Methods* (2nd ed.), Berlin: Springer. [1]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [4,5]

Shephard, N., and Pitt, M. K. (1997), "Likelihood Analysis of Non-Gaussian Measurement Time Series," *Biometrika*, 84, 653–667. [4,5]

Spantini, A., D. Bigoni, and Y. Marzouk (2018), "Inference via Low-Dimensional Couplings," *Journal of Machine Learning Research* 19, 1–71. [13]

Stan Development Team (2018), *RStan: the R interface to Stan* (Version 2.17.3). Available at: http://mc-stan.org. [7]

——— (2019), *Stan User's Guide* (version 2.20). Available at: http://mc-stan.org. [2,5]

Taylor, S. J. (1986), *Modelling Financial Time Series*, Chichester: Wiley. [8]

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86. [5]

Zhang, Y., and Sutton, C. (2014), "Semi-Separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models," in *Advances in Neural Information Processing Systems*, eds. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Vol. 27). Red Hook, NY: Curran Associates, Inc., pp. 10–18. [1]