Differentiable Bayesian inference of SDE parameters using a pathwise series expansion of Brownian motion

Sanmitra Ghosh[1] **Paul J. Birrell**[2,1] **Daniela De Angelis**[1,2]
{sanmitra.ghosh,paul.birrell,daniela.deangelis}@mrc-bsu.cam.ac.uk
[1]MRC Biostatistics Unit, University of Cambridge, [2]UK Health Security Agency

Abstract

By invoking a pathwise series expansion of Brownian motion, we propose to approximate a stochastic differential equation (SDE) with an ordinary differential equation (ODE). This allows us to reformulate Bayesian inference for a SDE as the parameter estimation task for an ODE. Unlike a nonlinear SDE, the likelihood for an ODE model is tractable and its gradient can be obtained using adjoint sensitivity analysis. This reformulation allows us to use an efficient sampler, such as NUTS, that rely on the gradient of the log posterior. Applying the reparameterisation trick, variational inference can also be used for the same estimation task. We illustrate the proposed method on a variety of SDE models. We obtain similar parameter estimates when compared to data augmentation techniques.

1 INTRODUCTION

The task of estimating the parameters of a SDE observed at discrete times is highly challenging. One has to deal with the estimation of a high dimensional latent diffusion in addition to the governing parameters. Moreover, the exact transition densities given by the forward Kolmogorov equation, required to calculate the likelihood, are intractable for a nonlinear SDE. Thus, the estimation task involves an intractable likelihood. This intractability can be side-stepped using the Euler-Maruyama discretisation of the SDE, resulting in a Gaussian approximation of the transition density. However, this approximation is valid for a small time interval. Often much smaller than the interval between successive observations. Hence, such approximation is generally used with *data-augmentation* methods (Sørensen, 2004) in the context of Bayesian inference using a Markov chain Monte Carlo (MCMC) algorithm. Additionally any sampling scheme has to also deal with the strong correlation between the diffusion path and the model parameters. Slow convergence of MCMC, that involves *data-augmentation*, incurs a high computational cost due to the iterative nature of sampling a discretised diffusion path.

With the widespread availability of automatic differentiation (AD) techniques, Bayesian inference is increasingly carried out using MCMC algorithms that traverse the parameter space based on gradient of the target density. Such algorithms have a proposal generating mechanism that can rapidly explore the parameter space when compared to traditional random-walk MCMC. Moreover, optimisation based alternatives to MCMC such as black-box variational inference has the potential to further expedite the inference process. AD has made it possible to apply such algorithms to many complex models with a differentiable target density. Many probabilistic programming platforms, that rely on AD, includes such algorithms and have vastly automated the entire process of Bayesian inference.

Application of the aforementioned efficient algorithms to a SDE is highly non-trivial. The likelihood is intractable and thus non-differentiable. If an Euler-Maruyama discretisation is used then one has to backpropagate through the SDE solver steps (Giles and Glasserman, 2006). This is error-prone and highly inefficient in terms of memory. Moreover, data-augmentation methods are often embedded within simulation based inference algorithms (Andrieu et al., 2010) that are inherently non-differentiable. Recently, an elegant method for extending AD to Stratonovich SDEs was introduced in Li et al. (2020) that is much more efficient (albeit for diagonal noise models) than earlier attempts. Building on the path-integral formulation of variational inference (Archambeau et al., 2008; Opper, 2019) to avoid the intractability of the transition

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

density, this AD technique can be used to carry out Bayesian inference of the diffusion path. However, this particular variational formulation does not lend itself to Bayesian inference of the parameters. Furthermore, this AD technique cannot be used in the context of MCMC, again due to the intractability of the transition density and thus the likelihood.

In case of an ODE, in contrast to a SDE, the likelihood is tractable and there exists an efficient numerical sensitivity analysis technique that lends itself to AD (Chen et al., 2018; Ghosh et al., 2021). Thus, we propose to approximate a SDE by an equivalent ODE using a pathwise truncated series expansion of Brownian motion (Lyons et al., 2012, 2014). The resulting ODE contains the same SDE parameters in addition to auxiliary parameters that are the expansion coefficients. We estimate all these parameters jointly where the marginal density of the model parameters is the desired quantity of interest. We carry out this estimation task using the No-U-Turn (Hoffman and Gelman, 2014) sampler (NUTS) as well as using black-box variational inference, both inherently capable of handling the inflated parameter dimension efficiently. We summarise our contributions as follows:

- We propose a method for estimating the parameters of a SDE by utilising an approximation scheme that transforms the SDE to an ODE. This approximation scheme enables the usage of efficient inference algorithms that use the gradient of the posterior distribution in order to traverse the parameter space.
- In contrast to some of the widely used methods for the inference of SDE parameters, the proposed method circumvents the need of data-augmentation which incurs a large computational expense.
- Application of the proposed method is not limited to SDEs that have a particular structure such as linear drift, constant diffusion and additive noise.

1.1 Related Work

Series expansion of Brownian motion for inference, in conjunction with a Gibbs sampling scheme, was first introduced in Lyons et al. (2012) for systems with additive noise. Lyons et al. (2014) applied series expansion for filtering in nonlinear state-space models. By marrying series expansion with adjoint sensitivity analysis, our approach can be applied to both additive and multiplicative noise models, the latter is prevalent in biology. Moreover, our approach is targeted towards differentiable inference, which accommodates faster and more efficient inference algorithms.

Pathwise approximation of SDE can be framed using the theory of *rough paths* (Friz and Hairer, 2020), going beyond Brownian motion. Our approach can be easily integrated with rough path based analysis. Pathwise approximations have also been used earlier (Kloeden and Jentzen, 2007) for numerical solution of SDEs, and recently for constructing normalizing flows (Hodgkinson et al., 2020). Our contribution is in using pathwise approximation to obtain a tractable and differentiable likelihood.

Variational inference for a SDE was first introduced in Archambeau et al. (2008) using a path-integral formulation. Li et al. (2020) extended these earlier approaches (Archambeau et al., 2008; Opper, 2019) using a novel AD technique. In addition to the aforementioned limitations of the method in Li et al. (2020), a common difficulty with all these path-integral approaches is that of finding a suitable variational approximation of the latent diffusion path. In our approach the variational approximation is constructed in the parameter space; a simpler task in comparison. We demonstrate this in section 6.3.

ODE approximation of transition density can also be derived using a Gaussian process approximation of the transition density (Golightly and Gillespie, 2013; Fearnhead et al., 2014). Our approach rely on a pathwise approximation of the diffusion and thus can accommodate non-Gaussian transition densities.

An AD framework for ODEs was introduced in Chen et al. (2018) utilising the adjoint sensitivity method of Pontryagin et al. (1962) and its application for inference (Stapor et al., 2018; Melicher et al., 2017).

2 BACKGROUND

We begin by first introducing the Bayesian inference framework for a diffusion process described by a SDE. Consider a K-dimensional diffusion process that satisfies the following SDE:

$$d\boldsymbol{X}_t = \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\theta})dt + \sqrt{\boldsymbol{B}(\boldsymbol{X}_t, \boldsymbol{\theta})}d\boldsymbol{W}_t, \quad \boldsymbol{X}_0 = \boldsymbol{x}_0, \tag{1}$$

where $\boldsymbol{X}_t$ denotes the value of the process at time t, $\boldsymbol{a}$ is a K-dimensional *drift* vector, $\boldsymbol{B}$ a $K \times K$ diffusion matrix and the driving noise $\boldsymbol{W}_t$ is a K-dimensional Brownian motion which is treated in the Itô sense. Both the drift and diffusion matrix depends on an unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$. If the initial value $\boldsymbol{x}_0$ is unknown then this is estimated along with the $\boldsymbol{\theta}$.

We assume that $\boldsymbol{a}(\cdot)$ and $\boldsymbol{B}(\cdot)$ are sufficiently regular functions such that Equation (1) has a weak non-explosive solution (Oksendal, 2013).

Consider a set of noisy experimental observations $\boldsymbol{y} \in \mathbb{R}^{M \times K}$ observed at M experimental time points, $\{t_i\}_{i=0}^{M-1}$, for the K states. Within the Bayesian inferential paradigm we want to place a prior distribution on the unknown parameters $p(\boldsymbol{\theta})$ and intend to obtain the corresponding posterior distribution

$$p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2}$$

of the latent path $\boldsymbol{X}$ and the unknown parameters $\boldsymbol{\theta}$, where $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ is the likelihood and $p(\boldsymbol{X}|\boldsymbol{\theta})$ is the distribution of the diffusion $\boldsymbol{X}$ defined from the SDE given by Equation (1). For a nonlinear SDE, $p(\boldsymbol{X}|\boldsymbol{\theta})$ is intractable. To work around this intractability an Euler-Maruyama discretisation is generally used, yielding a Gaussian approximation to the transition density. Inference proceeds by generating samples of $\boldsymbol{X}$ and $\boldsymbol{\theta}$.

Within an MCMC scheme, samples of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ can be drawn by using a Gibbs sampling (Golightly and Wilkinson, 2008) scheme. As an alternative to such Gibbs proposal mechanism both $\boldsymbol{X}$ and $\boldsymbol{\theta}$ can be jointly updated using a particle MCMC scheme wherein a particle filter is used to draw the diffusion path and evaluate an unbiased estimate of the likelihood. Particle MCMC methods, despite their higher computational cost, produce a faster mixing Markov chain and thus have become the state-of-the-art for inference in SDEs. Nonetheless, both these methods are inherently non-differentiable and rely crucially on a random-walk proposal for updating $\boldsymbol{\theta}$.

3 PATHWISE APPROXIMATION

By the definition of an Itô integral, within a time interval $[0, T]$ a standard Brownian motion can be written as (Luo, 2006; Lyons et al., 2012):

$$W_t = \int_0^t dW_s = \int_0^T \mathbb{I}_{[0,t]}(s) dW_s, \tag{3}$$

where $\mathbb{I}_{[0,t]}(\cdot)$ is the indicator function. Suppose $\{\phi_i\}_{i \geq 1}$ is an orthonormal basis of $L^2[0, T]$. For example this can be the following trigonometric function related to the Karhunen–Loève (KL) expansion of Brownian motion (Särkkä and Solin, 2019):

$$\phi_i(t) = (2/T)^{1/2} \cos\{(2i - 1)\pi t/2T\}. \tag{4}$$

We can interpret $\mathbb{I}_{[0,t]}$ as an element of $L^2[0, T]$, and expand it in terms of the basis functions:

$$\begin{aligned} \mathbb{I}_{[0,t]}(s) &= \sum_{i=1}^{\infty} \left\langle \mathbb{I}_{[0,t]}(\cdot), \phi_i(\cdot) \right\rangle \phi_i(s) \\ &= \sum_{i=1}^{\infty} \Big(\int_0^t \phi_i(u) du \Big) \phi_i(s). \end{aligned} \tag{5}$$

Substituting (5) into (3) we see that:

$$W_t = \sum_{i=1}^{\infty} \Big(\int_0^T \phi_i(s) dW_s \Big) \int_0^t \phi_i(u) du. \tag{6}$$

We will use the shorthand $Z_i := \int_0^T \phi_i(s) dW_s$. Since the basis functions $\{\phi_i\}$ are deterministic and orthonormal, it follows from standard results of Itô calculus that $Z_i \sim \mathcal{N}(0, 1)$.

The infinite series in Equation (6) can be truncated after N terms to obtain a pathwise series approximation $\hat{W}_t$ of Brownian motion. Taking derivative with respect to time we obtain the following approximation to the differential of Brownian motion given by

$$d\hat{W}_t = \sum_{i=1}^{N} Z_i \phi_i(t) dt. \tag{7}$$

Consider now a scalar SDE driven by this approximate Brownian motion:

$$dX_t = a(X_t, \theta)dt + b(X_t, \theta)d\hat{W}_t, \tag{8}$$

where a, b are some scalar drift and diffusion term. As $N \to \infty$ the exact solution of the equation above will converge to the solution had it been driven by the exact Brownian motion. The seminal work of Wong and Zakai (1965) shows that as $N \to \infty$ the solution however converges to the solution of a Stratonovich SDE given by

$$dX_t = a(X_t, \theta)dt + b(X_t, \theta) \circ dW_t, \tag{9}$$

Thus, we can approximate the above Stratonovich SDE by Equation (8), which is actually an ODE:

$$\frac{d\hat{X}_t}{dt} = a(\hat{X}_t, \theta) + b(\hat{X}_t, \theta) \sum_{i=1}^{N} Z_i \phi_i(t). \tag{10}$$

Similarly, we can approximate a K-dimensional Brownian motion by applying the truncated series expansion to each of the K dimensions. Substituting such a K-dimensional approximation in a Stratonovich equivalent of Equation (1) we obtain the following ODE:

$$\frac{d\hat{\boldsymbol{X}}_t}{dt} = \tilde{\boldsymbol{a}}(\hat{\boldsymbol{X}}_t, \boldsymbol{\theta}) + \sqrt{\boldsymbol{B}(\hat{\boldsymbol{X}}_t, \boldsymbol{\theta})} \sum_{i=1}^{N} \boldsymbol{Z}_i \Phi_i(t), \tag{11}$$

where $\boldsymbol{Z}_i, \Phi_i \in \mathbb{R}^K$ and $\tilde{\boldsymbol{a}}(\cdot)$ is the drift of the Stratonovich equivalent of Equation (1). Formula of $\tilde{\boldsymbol{a}}(\cdot)$ in terms of the Itô SDE's drift and diffusion is given in Appendix A. Note that the drift of the Itô and Stratonovich equivalents are same when $\boldsymbol{B}(\cdot)$ is not a function of the state $\boldsymbol{X}_t$.

Convergence to the Stratonovich SDE in the multivariate case is not guaranteed in general (Lyons et al., 2014). However, theorem 2.3.1 in Shmatkov (2006) states that convergence for the multivariate case is guaranteed when the sequence in Equation (6) converges uniformly. If one chooses to expand the Brownian motion using Haar wavelets then convergence of this sequence is indeed uniform (Lyons et al., 2014; McShane, 2020). This is the Levy-Ciecelski construction of Brownian motion. Note that in our experiments we did not observe a failure to converge to the Stratonovich limit while using the KL basis function. Similar findings were reported in Lyons et al. (2012).

Only a few terms in the series expansion in general captures the large scale oscillations whereas the remaining terms determine the small-scale (high frequency) oscillations. Thus, even without using a really large number of terms in the expansion a good approximation $p(\hat{\boldsymbol{X}}_t)$ to the true time t marginal distribution $p(\boldsymbol{X}_t)$ can be achieved. In Figure 1 we compare the marginal density of the states of a Lotka-Volterra model (see section 6.1) obtained through the Euler-Maruyama and the series approximation introduced here. Clearly, with even $N = 10$ the approximation matches well.

4 INFERENCE USING MCMC

Note that the randomness in the model in Equation (11) is now encapsulated in the expansion coefficients $\boldsymbol{Z}_i$. Consequently, inference involving the above model is no longer a "missing-data" problem that is generally solved using data-augmentation. Thus, we use the ODE approximation of the SDE for estimating the SDE parameters $\boldsymbol{\theta}$. We first assume that $\boldsymbol{y}$ is the vector of noisy observation, at the M time-points, of the solution $\hat{\boldsymbol{X}}(\boldsymbol{\theta}) := \hat{\boldsymbol{X}}(\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{x}_0) \in \mathbb{R}^K$ of the ODE given by Equation (11), which approximates the actual diffusion path $\hat{\boldsymbol{X}} \approx \boldsymbol{X}$, where $\mathbf{Z} := \{\boldsymbol{Z}_i\}_{i=1}^N$ and $\boldsymbol{x}_0$ is the initial value. The task then is to infer the posterior distribution of all the unknown parameters of the ODE: $\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{x}_0$. We can now write the joint posterior distribution of $(\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{x}_0)$ as

$$p(\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{x}_0|\boldsymbol{y}) \propto p(\boldsymbol{y}|\hat{\boldsymbol{X}}(\boldsymbol{\theta}))p(\mathbf{Z})p(\boldsymbol{\theta}, \boldsymbol{x}_0), \tag{12}$$

where the likelihood $p(\boldsymbol{y}|\hat{\boldsymbol{X}}(\boldsymbol{\theta}))$ is now both tractable and differentiable, with respect to $\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{x}_0$. Note that the prior $p(\boldsymbol{Z}_i)$ is $\mathcal{N}(\mathbf{0}, \mathbb{I}_{K\times K})$ by construction.

We can obtain samples from this posterior distribution using a MCMC, sampler that targets the gradient of the log posterior: $\nabla_{\boldsymbol{\theta},\mathbf{Z},\boldsymbol{x}_0}\{\log p(\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{x}_0|\boldsymbol{y})\}$, such as the NUTS sampler. The marginal density $p(\boldsymbol{\theta}|\boldsymbol{y})$ can be easily obtained by collecting the corresponding samples. Moreover, we can obtain the posterior distribution of the approximate sample path $\hat{\boldsymbol{X}}(\boldsymbol{\theta})|_{\boldsymbol{\theta},\mathbf{Z},\boldsymbol{x}_0\sim p(\boldsymbol{\theta},\mathbf{Z},\boldsymbol{x}_0|\boldsymbol{y})}$ by solving the ODE using the parameter samples.

5 VARIATIONAL INFERENCE

Let us denote by $\boldsymbol{\Theta} := (\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{x}_0)$ the vector containing all the unknown quantities that we want to estimate. Using variational inference we can approximate $p(\boldsymbol{\Theta}|\boldsymbol{y})$ with a tractable distribution $q(\boldsymbol{\Theta}|\boldsymbol{\lambda})$ from a family of distributions $q(\cdot|\boldsymbol{\lambda})$, indexed by $\boldsymbol{\lambda}$, by maximising the *evidence lower bound* (ELBO) (Jordan et al., 1999) given by

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}[\log p(\boldsymbol{y}|\hat{\boldsymbol{X}}(\boldsymbol{\Theta}))p(\boldsymbol{\Theta})] - \mathbb{E}[\log q(\boldsymbol{\Theta}|\boldsymbol{\lambda})] \tag{13}$$

where the above expectations are with respect to $q(\boldsymbol{\Theta}|\boldsymbol{\lambda})$. If gradient of the ELBO, w.r.t the variational parameter $\boldsymbol{\lambda}$, is available then variational inference can be formulated as a simple gradient descent problem as follows:

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \gamma\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}), \tag{14}$$

where γ is a learning rate. However, for an ODE the above expectations are intractable. Following Ghosh et al. (2021), we apply the reparameterisation trick (Kingma et al., 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014) to obtain a Monte Carlo (MC) estimate, using L samples, of the gradient of the ELBO:

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{MC}(\boldsymbol{\lambda}) = \frac{1}{L}\sum_{l=1}^{L}\nabla_{\boldsymbol{\Theta}}\Big[\log p(\boldsymbol{y}|\hat{\boldsymbol{X}}(\boldsymbol{\Theta}^l))p(\boldsymbol{\Theta}^l) - \log q(\boldsymbol{\Theta}^l|\boldsymbol{\lambda})\Big]\nabla_{\boldsymbol{\lambda}}g(\boldsymbol{\lambda}, \boldsymbol{\epsilon}^{(l)}), \tag{15}$$

where $\boldsymbol{\Theta}^l$ is the output of an invertible, differentiable function $g(\boldsymbol{\lambda}, \boldsymbol{\epsilon}^{(l)})$ and $\boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$–a parameter free distribution. Substituting this MC estimate in Equation (14), we can find the optimal $\boldsymbol{\lambda}^*$ using the following stochastic optimisation update:

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \gamma\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{MC}(\boldsymbol{\lambda}). \tag{16}$$

5.1 Choice of The Approximation

To place prior distributions with support on positive reals we need to transform the support of $\boldsymbol{\Theta}$ to the unconstrained real line $\mathbb{R}^D$: $T : \mathbb{R}^D_{>0} \to \mathbb{R}^D$, and subsequently obtain a transformed parameter vector $\boldsymbol{\xi} = T(\boldsymbol{\Theta})$. The posterior density $p(\boldsymbol{\Theta}|\boldsymbol{y})$, with the above transformation, is given by

$$p(\boldsymbol{\Theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\hat{\boldsymbol{X}}(T^{-1}(\boldsymbol{\xi})))p(T^{-1}(\boldsymbol{\xi}))\ \big|\det J_{T^{-1}}(\boldsymbol{\xi})\ \big|, \tag{17}$$

where $J_{T^{-1}}(\boldsymbol{\xi})$ is the Jacobian of the inverse of T. This transformation lets us choose an approximating distribution $q(\boldsymbol{\xi}|\boldsymbol{\lambda})$ with unconstrained support, such as a Gaussian.

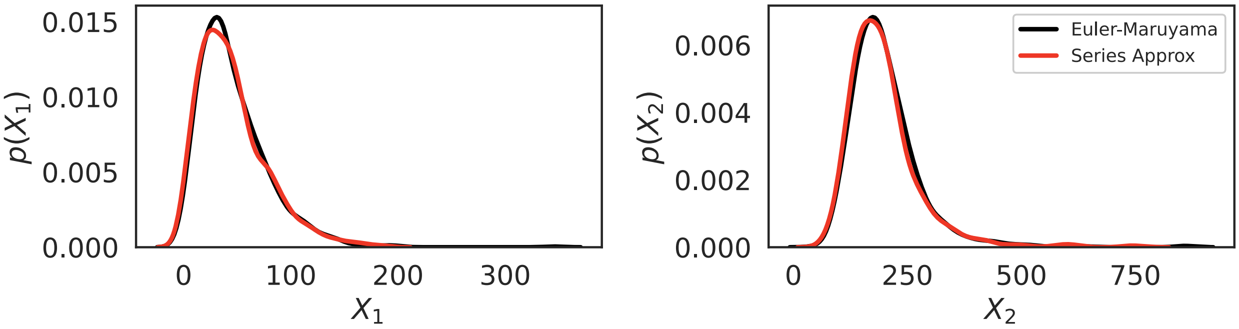


Figure 1: Comparison of the marginal density of the Lotka-Volterra model states at time T=30. We used 1000 samples to evaluate these densities.

Gaussian Approximation: Parameters of a nonlinear ODE such as Equation (11) are often strongly correlated. Moreover due the nature of our approximation we also expect correlations to exist between the expansion coefficients $\mathbf{Z}$ and $\mathbf{\Theta}$. For nonlinear ODEs, a full-rank Gaussian density as the variational approximation was shown in Ghosh et al. (2021) to be able to capture the correlation structure of the posterior amply. Thus, folowing Ghosh et al. (2021) we also chose a full-rank Gaussian approximation: $q(\boldsymbol{\xi}|\boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\lambda} = \big(\boldsymbol{\mu}, \mathbf{\Sigma}\big)$ are the variational parameters. To ensure that the covariance matrix $\mathbf{\Sigma}$ remains positive semidefinite, we parameterise the covariance using Cholesky factorisation, $\mathbf{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{'}$. To ensure uniqueness we take the logarithm of the diagonal elements of $\boldsymbol{L}$. The required reparameterisation, $\boldsymbol{\xi} = g(\boldsymbol{\lambda}, \boldsymbol{\epsilon})$, then simply follows as the affine transform $\boldsymbol{\xi} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$.

5.2 Gradient Evaluations

To carry out inference we need to obtain the gradient of a scalar function: i) the log posterior in the MCMC case or, ii) the ELBO for variatonal inference with respect to the model parameters $\mathbf{\Theta}$ or the variational parameters $\boldsymbol{\lambda}$. This in turn requires the propagation of gradients through the ODE. This can be efficiently carried out using the *adjoint sensitivity* analysis (Chen et al., 2018; Rackauckas et al., 2018; Ghosh et al., 2021), which is the continuous formulation of reverse-mode automatic differentiation. We describe this for obtaining the gradient of the ELBO w.r.t $\mathbf{\Theta}$. Note that the downstream gradients, $\frac{d\mathbf{\Theta}}{d\boldsymbol{\lambda}}$, can be trivially obtained using AD. Gradient of the prior and likelihood densities can be obtained analogously.

Consider a cost function, such as the ELBO, that depends on the ODE solution at the measurement times: $C(\hat{\boldsymbol{X}}) = \sum_i c(\hat{\boldsymbol{X}}_{t_i})$, where the sum appears due the factorisation of the likelihood over the time axis. In *adjoint sensitivity* analysis (Rackauckas et al., 2018) the gradient of the above scalar-valued cost function $C(\cdot)$, whose input is the ODE solution, can be computed directly. The first step is to solve a backwards ODE, the adjoint problem:

$$\frac{d\boldsymbol{a}(t)}{dt} = -\boldsymbol{a}(t)^{\mathrm{T}} \frac{\partial \boldsymbol{f}}{\partial \hat{\boldsymbol{X}}}, \tag{18}$$

where we use the shorthand $\boldsymbol{f}$ to denote denote the velocity field of the ODE in Equation (11). Furthermore, at each experimental time point t_i this backward ODE is perturbed by $\frac{\partial c(\hat{\boldsymbol{X}}_{t_i})}{\partial \hat{\boldsymbol{X}}}$. The gradient of the cost function with respect to the ODE parameters can be evaluated by another quadrature as follows:

$$\frac{dC}{d\boldsymbol{\theta}} = \boldsymbol{a}(t_0)^{\mathrm{T}} \frac{\partial \boldsymbol{f}\big(\hat{\boldsymbol{X}}_{t_0}, \mathbf{\Theta}\big)}{\partial \mathbf{\Theta}} + \sum_i \int_{t_i}^{t_{i+1}} \boldsymbol{a}(t)^{\mathrm{T}} \frac{\partial \boldsymbol{f}}{\partial \mathbf{\Theta}} dt. \tag{19}$$

Note that a continuous solution of the system and the adjoint states is required for the integration above. Alternatively, the system, the adjoint and the cost function ODEs can be solved simultaneously backward in time (Chen et al., 2018).

6 BENCHMARKING

We begin by testing the efficacy of the series approximation for the task of parameter estimation. We conducted two sets of experiments. In the first set we compared the proposed MCMC and variational inference schemes with the particle marginal Metropolis-Hastings (PMMH) algorithm (Andrieu et al., 2010; Golightly and Wilkinson, 2011), whose estimates of $\boldsymbol{\theta}, \boldsymbol{X}$ was treated as *gold standard* and a proxy for the true unknown posterior. In the second set of experiments we compared performance of the the proposed variational formulation with the one proposed in Li et al. (2020).

For the first set of experiments we used two biological SDEs: (i) the stochastic Lotka-Volterra model, and (ii) the stochastic SIR epidemic model. Both these models have non-diagonal diffusion and thus is less amenable to the method of Li et al. (2020) (see section 3.3 in that paper). Also, these models have state dependent diffusion term and thus their drifts were converted to the Stratonovich equivalents. For the second set of experiments we chose (i) the SDE associated with the Ornstein-Uhlenbeck (OU) process, and (ii) a SDE describing a bi-stable double-well system whose marginal density is multi-modal. We used simulated data for all the models except the SIR one.

While applying the PMMH algorithm, we used a Bootstrap particle filter (Gordon et al., 1995; Golightly and Wilkinson, 2011) which uses the Euler-Maruyama discretisation as the proposal (Golightly and Wilkinson, 2011) for $\boldsymbol{X}$ and we updated $\boldsymbol{\theta}$ using an adaptive random-walk Metropolis-Hastings algorithm.

Having run the PMMH, we then compared with this gold standard the estimates of the posterior obtained by applying MCMC using the NUTS algorithm and variational inference, both using the proposed series approximation: which we denote as **SA-ODE**. By VI we denote the proposed variational inference using **SA-ODE**, and by pathVI we denote the variational inference method of Li et al. (2020). Further details of MCMC and VI settings, for each experiment, is given in Appendix D.

A crucial algorithmic hyperparameter for the proposed method is the number of basis functions N in the truncated series expansion. We chose N by comparing the time T marginal densities obtained by solving (i) the **SA-ODE** and (ii) the original SDE using Euler-Maruyama discretisation, based on some trial parameters sampled from the prior. We used the N for which we found reasonable agreement between the marginal densities. For all the models except the double-well this was $N \geq 8$. However, for the double-well, which has a multi-modal marginal density, $N \geq 50$ was needed. Thus, we set $N = 10$ for all the models except double-well, for the later we set $N = 50$. Furthermore, we carried out a sensitivity analysis of the posterior estimates of the Lotka-Volterra model to the choice of N. This sensitivity analysis is furnished in Appendix B.2. We defer further discussion on the implications of the choice of N till section 7.

For all the models we have used the KL basis function (see Equation 4) to implement the **SA-ODE**. In Appendix B.1 we have compared the posterior estimates for the Lotka-Volterra, as obtained by NUTS, between the KL and the Haar wavelet basis. The estimates were indistinguishable. However, the wavelet basis function requires stricter error tolerances for the ODE solver. For this reason we decided to use the KL basis function throughout. Note that the OU and double-well SDEs are univariate and thus convergence is guaranteed for both the KL and wavelet functions. Finally, we have set the value of T in Equation (4) to be the end-point of the chosen time interval for each of the models.

We used the `numpyro` probabilistic programming library to apply NUTS and variational inference for the ease of comparison. Furthermore, we used a `jax` implementation of the Dormand-Prince adaptive ODE solver for integrating the **SA-ODE**. This solver provides an implementation of adjoint sensitivity that is needed for applying NUTS and variational inference. For the PMMH algorithm we implemented a vectorised particle filter in `jax` and implemented the adaptive MCMC in `Python`. The code is available at `https://github.com/sg5g10/sdeinference`.

6.1 Stochastic Lotka-Volterra Model

The stochastic Lotka–Volterra model (Wilkinson, 2018) has been widely used for benchmarking (see Fearnhead et al. (2014); Giagos (2010)). This model describes a population comprising of two competing species: *predators* which die with rate c_2 and reproduce with rate c_1 by consuming prey, which in turn reproduce with rate c_3. This system can be defined using the following drift and diffusion terms (see (Wilkinson, 2018) for derivation):

$$\boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\theta}) = \begin{bmatrix} c_1 X_t^1 - c_2 X_t^1 X_t^2 \\ c_2 X_t^1 X_t^2 - c_3 X_t^2 \end{bmatrix},$$
$$\boldsymbol{B}(\boldsymbol{X}_t, \boldsymbol{\theta}) = \begin{bmatrix} c_1 X_t^1 + c_2 X_t^1 X_t^2 & -c_2 X_t^1 X_t^2 \\ -c_2 X_t^1 X_t^2 & c_3 X_t^2 + c_2 X_t^1 X_t^2 \end{bmatrix}, \tag{20}$$

where the state vector $\boldsymbol{X}_t = (X_t^1, X_t^2)$ denotes the prey and predator species respectively. The parameter vector is the rate constants: $\boldsymbol{\theta} = (c_1, c_2, c_3)$, and the task is to estimate these given a noise corrupted sample path from the above system. We generated such a sample path using the Euler-Maruyama discretisation, with initial values $\boldsymbol{x}_0 = (100., 100.)$, between the time interval $[0 : 0.5 : 50]$. A set of 10 evenly spaced values from this path corrupted with Gaussian noise with $\sigma = 10$ constitute the observations $\boldsymbol{y}$. Following Golightly and Wilkinson (2011) we consider $\boldsymbol{x}_0, \sigma$ to be known and thus we are left with estimating the rate constants and the expansion coefficients.

The likelihood, for the series approximation, is a Gaussian, $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{Z}, \sigma) = \prod_i \mathcal{N}(\hat{\boldsymbol{X}}(t_i; \boldsymbol{\theta}, \mathbf{Z}), \sigma^2 \mathbb{I})$, where $\mathbb{I}$ is a 2×2 identity matrix. We placed the following priors on the rates: $c_1 \sim \text{Beta}(2, 1)$ $c_2 \times 100 \sim \text{Half}\,\mathcal{N}(0, 1)$

and $c_3 \sim \text{Beta}(1,2)$. The priors on the expansion coefficients are just standard Gaussians.

Summaries of the posterior marginals are furnished in Table 1. Plots of the posterior diffusion paths are given in Appendix C.1. We noticed close match between the posteriors estimated with the series approximation to that of the PMMH. The estimates obtained using NUTS, however, are closer to the gold standard. NUTS produced a relative ESS (averaged over the three parameters) of 0.89 and PMMH produced 0.05. We repeated the inference process using further realisations of the artificial noise. Additional results are furnished in Appendix B.3.

Runtimes of the competing algorithms are furnished in Table 2. The large difference between the runtimes of PMMH and NUTS/VI is due to the additional computational expense incurred by the particle filter, in comparison to solving the **SA-ODE**, at each step of the MCMC. However, this additional computational expense can be reduced by implementing the particle filter on a GPU (or a high-performance computing cluster).

6.2 The SIR Compartmental Model

The SIR model (Anderson et al., 1992) of infectious disease models the number of susceptible (S), infected (I), and recovered (R) people in a population subjected to an epidemic. The stochastic version of the SIR model, for a population of N people, can be defined using an SDE with the following drift and diffusion terms (see Fuchs (2013) for derivation):

$$\begin{aligned} \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\theta}) &= \begin{bmatrix} -\beta S_t I_t \\ \beta S_t I_t - \gamma I_t \end{bmatrix}, \\ \boldsymbol{B}(\boldsymbol{X}_t, \boldsymbol{\theta}) &= \frac{1}{N} \begin{bmatrix} \beta S_t I_t & -\beta S_t I_t \\ -\beta S_t I_t & \beta S_t I_t + \gamma I_t \end{bmatrix}, \end{aligned} \tag{21}$$

where the infection β and recovery γ rates are unknown parameters. Also, we have $N = S_t + I_t + R_t$. We used this SDE to model an outbreak of influenza at a boys boarding school in 1978 (Jackson et al., 2013). This particular dataset was previously used for benchmarking in Ryder et al. (2018). This data consists of the number of infections for a period of 14 days. The population size is $N = 763$. In addition to β, γ we also estimated the fractional initial susceptibility, $s_0 = S_0/N$, assuming the initial recovered fraction $r_0 = 0$ and thus $i_0 = 1 - s_0$. As this is count data we have used a Poisson likelihood $p(y_t|\beta, \gamma, s_0, \mathbf{Z}) = \text{Poisson}(I_t)$, and placed the following priors: $\beta, \gamma \sim \text{Gamma}(2,2)$ and $s_0 \sim \text{Beta}(2,1)$.

Summaries of the posterior marginals are presented in Table 1. The model fit plot is shown in Appendix C.2. In this example NUTS produced a relative ESS (again averaged over the three parameters) of 0.95 and PMMH produced 0.05. Runtimes are furnished in Table 2, where we noticed similar speedup as was found in the case of Lotka-Volterra model.

6.3 Comparison with an Alternative Variational Formulation

For the purpose of comparing VI with the pathVI method we used the OU process given by the following SDE:

$$dx_t = \theta_1(\theta_2 - x_t)dt + \theta_3 dW_t, \tag{22}$$

with unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, and the double-well SDE given by

$$dx_t = 4x_t(\theta_1 - x_t^2)dt + \theta_2 dW_t, \tag{23}$$

with unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$. For both these SDEs we simulated noisy data by adding Gaussian noise with $\sigma = 0.05$ to the diffusion obtained between the time interval $[0 : 0.1 : 10]$ with initial value $x_0 = 0$. We repeated the above to generate 5 datasets for each model. We estimated x_0, σ for both models.

In Li et al. (2020) the pathVI algorithm was formulated to produce a point estimate of the parameters and an approximate posterior distribution of the latent sample path. Thus, we compared the accuracy of the point estimates of the parameters and posterior distribution of the sample path obtained using VI and pathVI. We also compared the performance of both these methods against PMMH. For VI and PMMH we used a MAP estimate of the parameters. For both the SDEs, while running PMMH and VI, we placed the following priors: $\boldsymbol{\theta} \sim \text{Gamma}(2,2)$, $x_0 \sim \mathcal{N}(0,1)$, $\sigma \sim \text{Half}\,\mathcal{N}(0,1)$, and we have a Gaussian likelihood, $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{Z}, x_0, \sigma) = \prod_i \mathcal{N}(x(t_i; \boldsymbol{\theta}, \boldsymbol{Z}), \sigma^2)$.

In pathVI one has to construct a variational approximation as a posterior SDE. This posterior SDE shares the same diffusion term with the prior SDE: which is the model whose parameters are to be inferred (as point estimates). However, one has to come up with a function describing the drift term in the variational approximation. Following Li et al. (2020) we chose the drift term as a mlp. Specifically we used a mlp with two hidden layers having 50 units each and a softplus nonlinearity. We tried some other combinations of layers/units/nonlinearity, but did not notice any significant change in performance. We used the implementation of pathVI found in the `torchsde`[1] package.

We compared the point estimates, obtained by the competing methods, by calculating the Euclidean distances between the ground truth $\boldsymbol{\theta}$ and its point estimate $\hat{\boldsymbol{\theta}}$. We summarised these distances across the five datasets, for both models, in Figure 3. We noticed that VI was

[1] `https://github.com/google-research/torchsde`

Table 1: The **mean ± standard deviation** of the posterior distribution of each parameter for the **Lotka-Volterra** and **SIR** models. **VI** and **NUTS** are using the **SA-ODE** model. Posterior distribution for each method is represented, pointwise, by 1000 samples.

The SIR model				
$\boldsymbol{\theta}$	True Value	**PMMH**	**VI**	**NUTS**
β	–	1.8427 ± 0.0719	1.8069 ± 0.1319	1.8479 ± 0.1413
γ	–	0.4875 ± 0.0190	0.4849 ± 0.0278	0.4851 ± 0.0258
s_0	–	0.9964 ± 0.0010	0.9957 ± 0.0010	0.9959 ± 0.0014

The Stochastic Lotka-Volterra model				
$\boldsymbol{c}$	True Value	**PMMH**	**VI**	**NUTS**
c_1	0.5	0.5081 ± 0.0261	0.4846 ± 0.0205	0.4961 ± 0.0219
$100 \times c_2$	0.25	0.2492 ± 0.0113	0.2417 ± 0.0092	0.2454 ± 0.0096
c_3	0.3	0.2965 ± 0.0159	0.2872 ± 0.0115	0.2924 ± 0.0126

able to produce more accurate estimates in comparison to pathVI. Moreover, VI's estimates were much more closer to the ones produced by PMMH. In Figure 2 we have shown the posterior distributions of the latent diffusion's sample path for two (out of the 5) datasets. It is evident from these plots that pathVI performs poorly in estimating the sample path of the latent diffusion. VI, in comparison, produced an estimate of the sample path that was closer to the gold standard produced by PMMH. The pathVI method minimises the KL divergence between the true and approximate posterior diffusion paths. Note that VI also minimises this KL divergence between the true and approximate posterior paths, implicitly, by minimising the divergence between the true and approximate posterior distribution of the coefficients. Constructing an adequate variational approximation of the posterior diffusion path, as is done in pathVI through the usage of a posterior SDE, is much more challenging than constructing a variational approximation of the posterior of the coefficients, which have a standard Gaussian prior. For this reason we noticed significantly better quality of estimates obtained using VI, especially for the double-well model which has a multi-modal time t marginal distribution $p(x_t)$.

Table 2: Runtimes of **PMMH**, **NUTS** and **VI** for the stochastic lotka-Volterra (**LV**) and **SIR** models. Respective parameter estimates are given in Table 1. These were run on a 3.6 GHz machine with 16 GB memory.

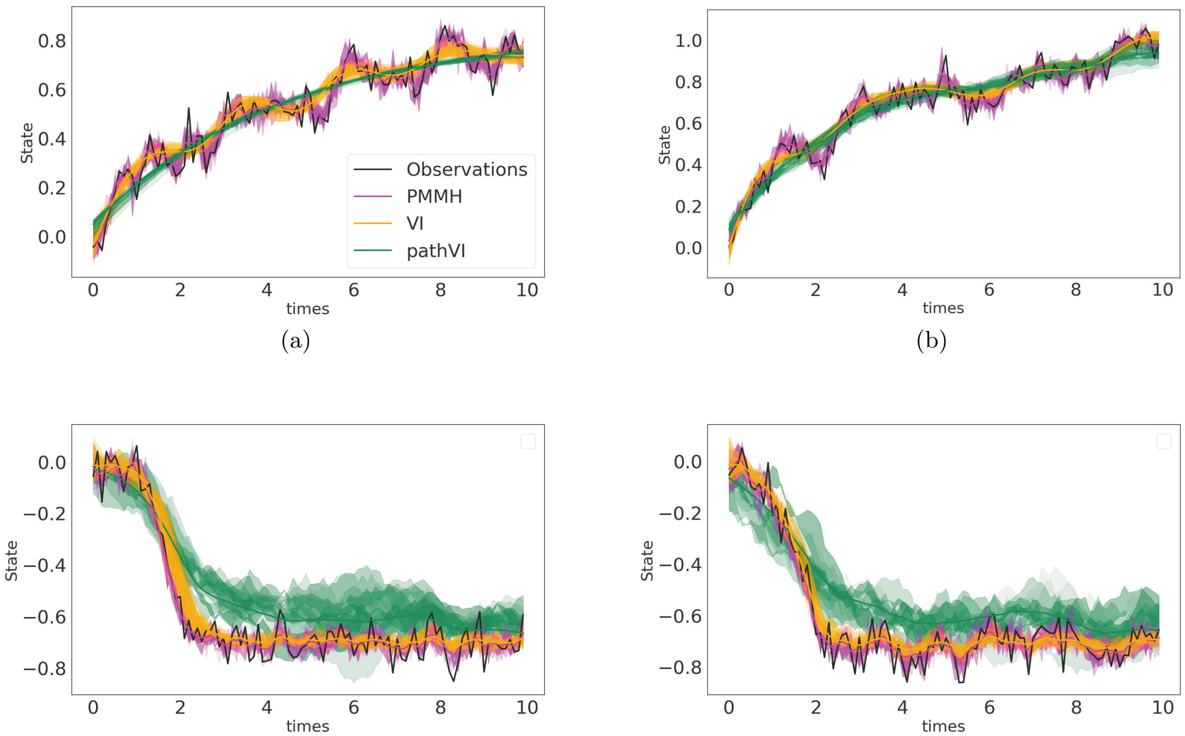
Runtimes in seconds			
Example	**PMMH**	**VI**	**NUTS**
LV	4919	75	203
SIR	3676	78	292

7 DISCUSSION

Clearly there is a trade-off between speed and accuracy of VI, in comparison to NUTS. However, this trade-off is not unique to this method and is a well known limitation of variational inference.

Applying the proposed method to high-dimensional SDEs may appear to be challenging considering the requirement of inferring a large number of coefficients. However, note that using any data-augmentation based inference method for SDEs, one has to infer the sample path of the latent diffusion, in addition to the model parameters. The sample path for a high-dimensional SDE scales much more poorly, in comparison to the coefficients, and inference is thus more challenging in comparison to the proposed method. The proposed method infers the (approximate) sample path through the coefficients.

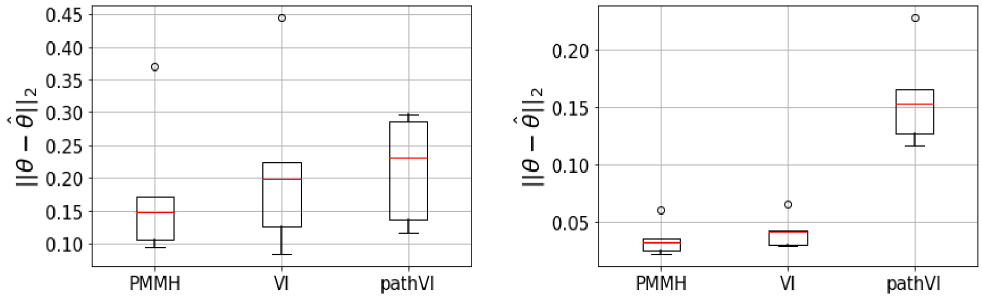
Theoretical bounds on the error between the posterior induced by a Gaussian process and its series approximation can be found in the PDE constrained inverse problem literature. See for example Dodwell et al. (2015) and the references therein. Similar bounds, with respect to N, can be derived for the proposed **SA-ODE** based inference. Alternatively, one can set a very large N in the **SA-ODE** and then approach the resulting high-dimensional sampling problem using a dimension-independent sampler (Cotter et al., 2013; Titsias and Papaspiliopoulos, 2018). Note that the complexity of **SA-ODE** does not depend on N by construction. Our simple recipe of comparing the time T marginal densities (see Figure 1) is found to be effective for the models we have used. We also noticed that a threshold value of N exists beyond which the increase in expansion terms has little effect on the estimation quality. This is reinforced through the sensitivity analysis (to the choice of N) we carried out for the Lotka-Volterra



(a) (b)

(c) (d)

Figure 2: Posterior distribution of the latent diffusion path, for two different datasets: (a-b) the **OU** process, (c-d) the **double-well** system. Summaries of the posterior distribution are shown in Appendix C.3.



(a) (b)

Figure 3: Euclidean distances between the ground truth and point estimates found by the three alternative methods for: (a) the **OU** process, (b) the **double-well** system. These distances are summarised across five datasets, for each model.

model (see Appendix B.2).

8 CONCLUSION

We presented a Bayesian inference approach for SDEs which replaces the SDE by a (random) ODE, rendering a tractable and differentiable likelihood. Using this approach we re-purposed differentiable inference algorithms for ODEs for the task of parameter estimation of SDEs. When compared to the particle MCMC algorithm, considered to be the state-of-the-art, we recovered similar posterior estimates of the parameters of a variety of SDEs. The proposed method also outperforms a recently proposed variational inference algorithm for SDEs, in regards to the accuracy of parameter estimates. In future work we want to develop the proposed methodology towards constructing generative models for time series data.

Acknowledgements

We like to thank the anonymous reviewers for their helpful comments and suggestions. SG was supported by the Medical Research Council (Unit programme number MC UU 00002/11).

References

Roy M Anderson, B Anderson, and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-taylor. Variational inference for diffusion processes. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. Mcmc methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

Tim J Dodwell, Christian Ketelsen, Robert Scheichl, and Aretha L Teckentrup. A hierarchical multilevel markov chain monte carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1075–1108, 2015.

Paul Fearnhead, Vasilieos Giagos, and Chris Sherlock. Inference for reaction networks using the linear noise approximation. *Biometrics*, 70(2):457–466, 2014.

Peter K Friz and Martin Hairer. *A course on rough paths*. Springer, 2020.

Christiane Fuchs. *Inference for diffusion processes: with applications in life sciences*. Springer Science & Business Media, 2013.

Sanmitra Ghosh, Paul Birrell, and Daniela De Angelis. Variational inference for nonlinear ordinary differential equations. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2719–2727. PMLR, 13–15 Apr 2021.

Vasileios Giagos. *Inference for auto-regulatory genetic networks using diffusion process approximations*. PhD thesis, Lancaster University, 2010.

Mike Giles and Paul Glasserman. Smoking adjoints: Fast monte carlo greeks. *Risk*, 19(1):88–92, 2006.

Andrew Golightly and Colin S Gillespie. Simulation of stochastic kinetic models. In *In Silico Systems Biology*, pages 169–187. Springer, 2013.

Andrew Golightly and Darren J Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693, 2008.

Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–820, 2011.

Neil Gordon, David Salmond, and Craig Ewing. Bayesian state estimation for tracking and guidance using the bootstrap filter. *Journal of Guidance, Control, and Dynamics*, 18(6):1434–1443, 1995.

Liam Hodgkinson, Chris van der Heide, Fred Roosta, and Michael W. Mahoney. Stochastic normalizing flows. *arXiv preprint arXiv:2002.09547v2*, 2020.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Charlotte Jackson, Emilia Vynnycky, Jeremy Hawker, Babatunde Olowokure, and Punam Mangtani. School closures and influenza: systematic review of epidemiological studies. *BMJ open*, 3(2), 2013.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

D P Kingma, Max Welling, et al. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 1, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Peter E Kloeden and Arnulf Jentzen. Pathwise convergent higher order numerical schemes for random ordinary differential equations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2087):2929–2944, 2007.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3870–3882. PMLR, 26–28 Aug 2020.

Wuan Luo. *Wiener chaos expansion and numerical solutions of stochastic partial differential equations.* California Institute of Technology, 2006.

Simon M. J. Lyons, Amos J. Storkey, and Simo Särkkä. The coloured noise expansion and parameter estimation of diffusion processes. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1961–1969, 2012.

Simon MJ Lyons, Simo Särkkä, and Amos J Storkey. Series expansion approximations of brownian motion for non-linear kalman filtering of diffusion processes. *IEEE Transactions on Signal Processing*, 62(6):1514–1524, 2014.

EJ McShane. Stochastic differential equations and models of random processes. In *Contributions to Probability Theory*, pages 263–294. University of California Press, 2020.

Valdemar Melicher, Tom Haber, and Wim Vanroose. Fast derivatives of likelihood functionals for ode based models using adjoint-state method. *Computational Statistics*, 32(4):1621–1643, 2017.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications.* Springer Science & Business Media, 2013.

Manfred Opper. Variational inference for stochastic differential equations. *Annalen der Physik*, 531(3): 1800233, 2019.

Liev Semiónovich Pontryagin, EF Mishchenko, VG Boltyanskii, and RV Gamkrelidze. *The Mathematical Theory of Optimal Processes.* 1962.

Christopher Rackauckas, Yingbo Ma, Vaibhav Dixit, Xingjian Guo, Mike Innes, Jarrett Revels, Joakim Nyberg, and Vijay Ivaturi. A comparison of automatic differentiation and continuous sensitivity analysis for derivatives of differential equation solutions. *arXiv preprint arXiv:1812.01892*, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR, 2014.

Tom Ryder, Andrew Golightly, A. Stephen McGough, and Dennis Prangle. Black-box variational inference for stochastic differential equations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4423–4432, 2018.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations.* Cambridge University Press, 2019.

Anton Shmatkov. *Rate of convergence of Wong-Zakai approximations for SDEs and SPDEs.* PhD thesis, The University of Edinburgh, 2006.

Helle Sørensen. Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354, 2004.

Paul Stapor, Fabian Fröhlich, and Jan Hasenauer. Optimization and profile calculation of ode models using second order adjoint sensitivity analysis. *Bioinformatics*, 34(13):i151–i159, 2018.

T. Tieleman and G Hinton. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning, 4*, 2016.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1971–1979. PMLR, 2014.

Michalis K. Titsias and Omiros Papaspiliopoulos. Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):749–767, 2018.

Darren J Wilkinson. *Stochastic modelling for systems biology.* CRC press, 2018.

Eugene Wong and Moshe Zakai. On the Convergence of Ordinary Integrals to Stochastic Integrals. *The Annals of Mathematical Statistics*, 36(5):1560 – 1564, 1965.

Supplementary Material: Differentiable Bayesian inference of SDE parameters using a pathwise series expansion of Brownian motion

A DRIFT OF STRATONOVICH SDE

The Stratonovich SDE given by

$$dX_t^i = \tilde{a}(X_t^i, \boldsymbol{\theta})dt + \sum_{j=1}^{K} \sqrt{b^{i,j}(X_t, \boldsymbol{\theta})} \circ dW_t^j, \quad i, j = 1, \dots K, \tag{24}$$

with the same solutions as the K-dimensional Ito SDE driven by a K-dimensional Wiener process given by

$$dX_t^i = a(X_t^i, \boldsymbol{\theta})dt + \sum_{j=1}^{K} \sqrt{b^{i,j}(X_t, \boldsymbol{\theta})} dW_t^j, \quad i, j = 1, \dots K, \tag{25}$$

has a drift coefficient that is defined component-wise as

$$\tilde{a}(X_t^i, \boldsymbol{\theta})dt = a(X_t^i, \boldsymbol{\theta})dt + \sum_{k=1}^{K} \sum_{j=1}^{K} \sqrt{b^{k,j}(X_t, \boldsymbol{\theta})} \frac{\partial \{\sqrt{b^{i,j}(X_t, \boldsymbol{\theta})}\}}{\partial X_t^k}. \tag{26}$$

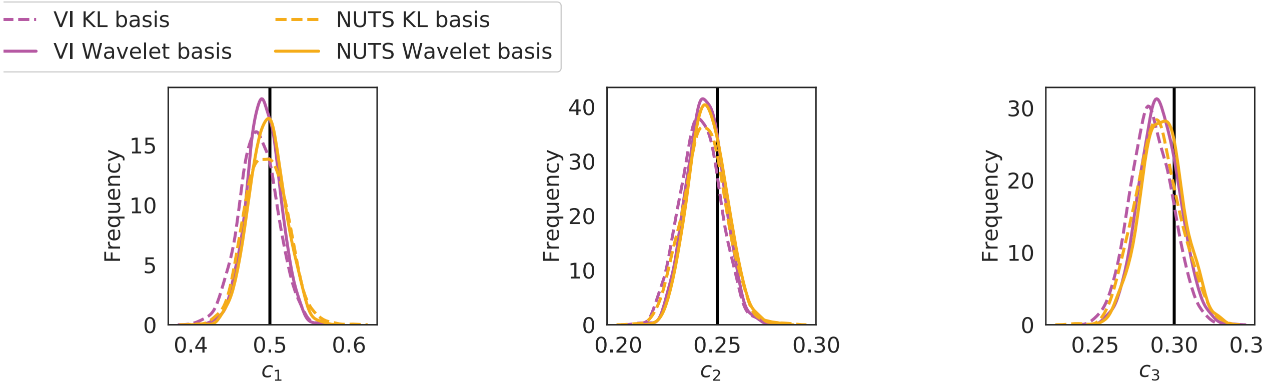


Figure 4: Comparison of the marginal density of the Lotka-Volterra parameters as obtained by NUTS and VI, using the wavelets (solid lines) and KL (dashed lines) basis function. The black vertical lines indicate the true parameter values.

B ADDITIONAL RESULTS FOR THE LOTKA-VOLTERRA MODEL

B.1 Wavelet Basis Function

For the multivariate case uniform convergence of the SA-ODE to the corresponding Stratonovich SDE is guaranteed if one chooses Haar wavelets as an orthonormal basis in which to expand the driving Brownian motion (Lyons et al., 2014). Thus, we wanted to compare the approximation achieved by using the KL basis function with that of Haar wavelets. But, before we delve into the comparison, let us briefly introduce the Haar wavelets: which is a complete orthonormal basis of $L^2[0,T]$.

The Haar wavelets are parameterised by two natural numbers: the scale $n \geq 0$ and the shift $0 \leq k < 2^n$. The first wavelet is defined as

$$\boldsymbol{\psi}_{0,0}(t) = \begin{cases} 1 & 0 \leq t < \frac{T}{2} \\ -1 & \frac{T}{2} < t \leq T \\ 0 & \text{otherwise} \end{cases}. \tag{27}$$

Further wavelets are defined by rescaling $\boldsymbol{\psi}_{0,0}$ so that it is non-zero only on some sub-interval of $[0,T]$, while ensuring that the wavelet still has unit norm. In general,

$$\boldsymbol{\psi}_{n,k}(t) = \frac{2^{n/2}}{\sqrt{T}}\boldsymbol{\psi}_{0,0}(2^n - kt). \tag{28}$$

Thus, $\boldsymbol{\psi}_{1,0}$ is a copy of $\boldsymbol{\psi}_{0,0}$ restricted to $[0,T/2]$, and $\boldsymbol{\psi}_{1,1}$ is a copy restricted to $[T/2,T]$. Furthermore, we add the constant function $\boldsymbol{\psi}_* = \frac{1}{\sqrt{T}}$ to form a complete basis. To be consistent with the notation introduced in section 3 (main text) we set $\boldsymbol{\phi}_1 = \boldsymbol{\psi}_*$, $\boldsymbol{\phi}_2 = \boldsymbol{\psi}_{0,0}$, $\boldsymbol{\phi}_3 = \boldsymbol{\psi}_{1,0}$ and so on.

To compare the wavelet and KL basis we ran NUTS and VI on the simulated dataset used in section 6.1 (main text), with the same algorithmic settings retained. We set $N = 10$ and $T = 50$ as was done in section 6.1 (main text). Marginal densities of the parameters are plotted in Figure 4. Although we get similar estimates, as in (Lyons et al., 2014, 2012), following which we used KL expansion throughout, the wavelets required stricter error tolerances for the ODE solver. We thus recommend the usage of a stiff solver when using the wavelet basis. Extension of AD for a stiff solver can be done using the custom op creation method of Ghosh et al. (2021). However, a JIT compiled solver, provided with `Jax`, is faster.

B.2 Sensitivity to Truncation of The Series: The Choice of N

To carry out this sensitivity analysis we used five realisations of the simulated dataset. The first one is used in section 6.1 (main text) and the estimates on the rest are summarised in section B.3. We ran the NUTS algorithm with $N = 3, 5, 8, 10$ respectively and measured the maximum mean discrepancy (MMD) to the corresponding estimates obtained by PMMH for each dataset. We used the KL basis throughout and retained all algorithmic settings for NUTS as in section 6.1 (main text). MMD for specific choices of N are shown in Figure 5. It is apparent that with $N = 8$ the MMD plateaus.

B.3 Results for Additional Datasets

To benchmark the methods with this model we used simulated data for which the corresponding estimates are summarised in Table 1 (main text). In addition to this dataset we generated four more datasets using new realisations of the artificial noise corruption. The MMD between the posterior distribution obtained using PMMH and the ones obtained using VI/NUTS are shown in Figure 6. All the algorithmic and model specific settings were kept the same as was used in section 6.1 (main text). Additionally in Table 3 we have summarised the individual posterior distributions for each of these additional dataset.

C ADDITIONAL PLOTS FOR EACH MODEL

C.1 Plots of The Latent Diffusion for Lotka-Volterra Model

Here we evaluate the posterior predictive distribution of the latent diffusion $\hat{\boldsymbol{X}}^*(\boldsymbol{\theta})|_{\boldsymbol{\theta},\mathbf{Z},\boldsymbol{x}_0\sim p(\boldsymbol{\theta},\mathbf{Z},\boldsymbol{x}_0|\boldsymbol{y})}$, using the **SA-ODE** approximation, on a finer grid t^* than the observations. The posterior predictive distribution is

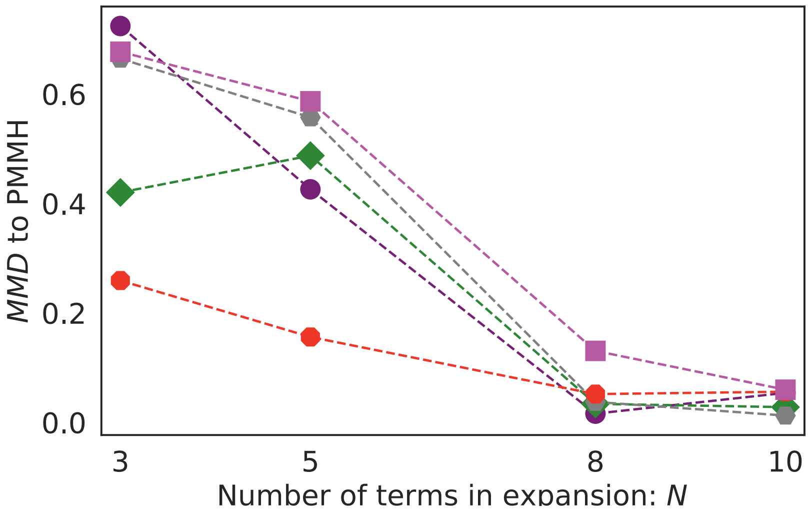


Figure 5: Comparison of the effect of increasing the number of expansion terms N on the MMD between the PMMH estimate and the estimates obtained from running NUTS, with $N = 3, 5, 8, 10$, for the Lotka-Volterra model. All estimates used 1000 samples from the posterior.

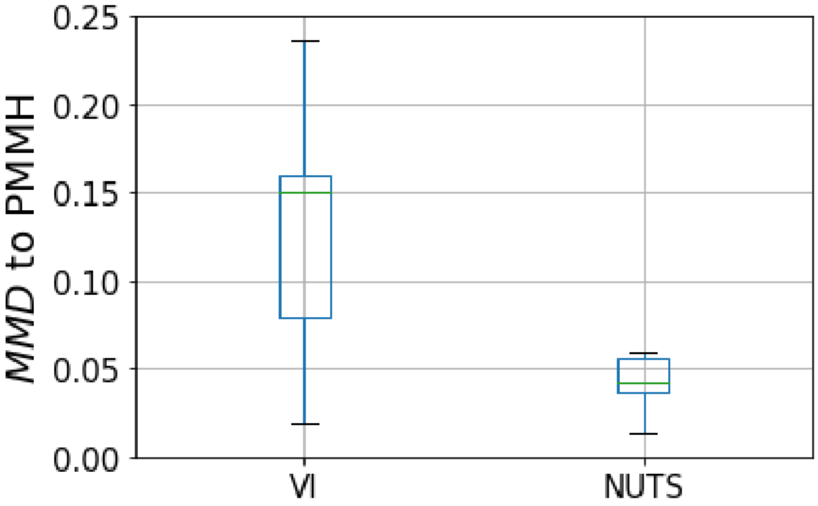


Figure 6: Comparison of the MMD between the posterior distribution obtained using PMMH to the ones obtained using NUTS/VI, for the stochastic Lotka-Volterra model. These MMDs are summarised across five datasets.

evaluated pointwise using samples obtained from running NUTS or drawn from the VI approximation. We used the posterior distributions estimated using the simulated dataset that was used in section 6.1 (main text). The mean and the 95% credible intervals of $\hat{\boldsymbol{X}}^*$ are shown in Figure 7.

C.2 Model Fit Plots for The SIR Model

We evaluated the posterior predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{y})$ on a finer time grid t^*, using the **SA-ODE**. Samples of the posterior predictive distribution were evaluated pointwise using the posterior estimates obtained from running NUTS and VI. Figure 8 summarises the mean and 95% credible intervals of $p(\boldsymbol{y}^*|\boldsymbol{y})$.

Table 3: The **mean ± standard deviation** of the posterior distribution of each parameter for the Stochastic Lotka-Volterra model, for each of the additional dataset.

		Dataset 2		
$\boldsymbol{c}$	True Value	**PMMH**	**VI**	**NUTS**
c_1	0.5	0.5194 ± 0.0273	0.5026 ± 0.0205	0.5093 ± 0.0224
$100 \times c_2$	0.25	0.2557 ± 0.0122	0.2477 ± 0.0092	0.2494 ± 0.0095
c_3	0.3	0.2937 ± 0.0163	0.2883 ± 0.0125	0.2906 ± 0.0132
		Dataset 3		
$\boldsymbol{c}$	True Value	**PMMH**	**VI**	**NUTS**
c_1	0.5	0.5172 ± 0.0213	0.4970 ± 0.0159	0.5210 ± 0.0204
$100 \times c_2$	0.25	0.2540 ± 0.0080	0.2432 ± 0.0067	0.2514 ± 0.0076
c_3	0.3	0.3140 ± 0.0117	0.3048 ± 0.0101	0.3130 ± 0.0111
		Dataset 4		
$\boldsymbol{c}$	True Value	**PMMH**	**VI**	**NUTS**
c_1	0.5	0.4696 ± 0.0200	0.4628 ± 0.0195	0.4574 ± 0.0205
$100 \times c_2$	0.25	0.2512 ± 0.0100	0.2499 ± 0.0077	0.2454 ± 0.0087
c_3	0.3	0.2836 ± 0.0134	0.2831 ± 0.0115	0.2778 ± 0.0127
		Dataset 5		
$\boldsymbol{c}$	True Value	**PMMH**	**VI**	**NUTS**
c_1	0.5	0.5129 ± 0.0197	0.5291 ± 0.0162	0.5259 ± 0.0216
$100 \times c_2$	0.25	0.2578 ± 0.0080	0.2646 ± 0.0067	0.2604 ± 0.0080
c_3	0.3	0.3229 ± 0.0123	0.3281 ± 0.0099	0.3215 ± 0.0109

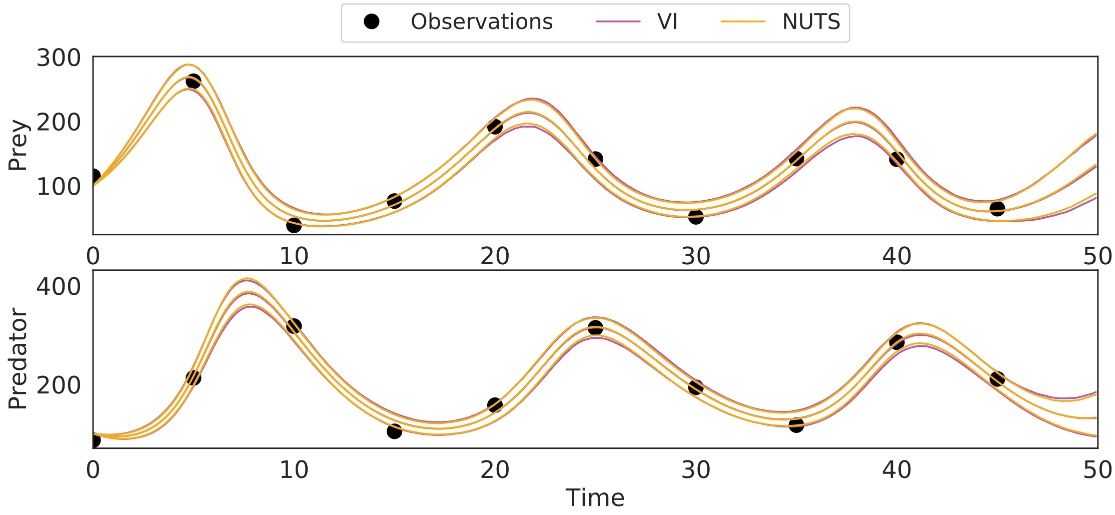


Figure 7: Mean and 95% credible intervals of the posterior predictive distribution of the latent diffusion $\hat{\boldsymbol{X}}^*$, for the Lotka-Volterra model.

C.3 Summaries of the Posterior Distribution of OU/DW Sample Paths

Summaries of the posterior distribution (shown in Figure 2(a) & 2(b) in the main text) of the latent sample path of the **OU** process, for two datasets, are shown in Figure 9. Summaries of the posterior distribution (see Figure 2(c) & 2(d) in the main text) of the sample path for the **double-well** model, for two datasets, are shown in Figure 10. These distributions were summarised by the mean and the 95% credible intervals.

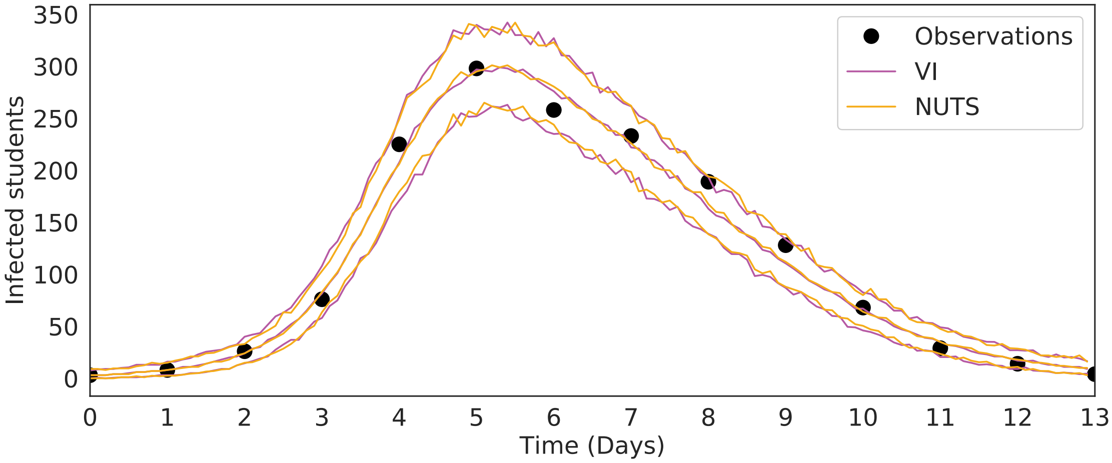


Figure 8: Mean and 95% credible intervals of the posterior predictive distribution for the SIR model.

D FURTHER DETAILS OF THE EXPERIMENTS

For, variational inference we used the RMSprop (Tieleman and Hinton, 2016) optimisation algorithm with a step size of 10^{-3}, for the first set of experiments, and 10^{-2} for the second set. Our choice of RMSprop is motivated by the findings in Ghosh et al. (2021). For the variational inference method of Li et al. (2020), we used the ADAM optimisation (Kingma and Ba, 2015) with a learning rate of 10^{-2} that is exponentially decayed with rate 0.999 during each iteration.

For the stochastic Lotka-Volterra model we ran two chains of PMMH, each for 100,000 iterations from slightly differing initial states. We ran the particle filter with 500 particles. We discarded the first 50,000 iterations as burnin for each chain and thinned accordingly to have 1000 samples representing the gold standard posterior estimate. For the **SA-ODE** two chains of NUTS were run for 1000 iterations (which is sufficient since NUTS is a high ESS sampler) after an initial 1000 warmup iterations. The NUTS samples, from the two chains, are then thinned to obtain 1000 samples. VI with the series approximation was run for 30,000 iterations with $L = 1$. For plotting and summarising the posterior distributions, and comparing to the gold standard we used 1000 samples from the variational approximation for this and the subsequent (SIR model) example.

For the stochastic SIR model we ran two chains of PMMH, each for 200,000 iterations, again started with slightly different initial states. We ran the particle filter with 500 particles. In this case we discarded the first 100,000 iterations as burnin for each chain and thinned accordingly to obtain 1000 samples that represent the gold standard. We used the same setup, as was used in the previous example, for applying NUTS and VI with the series approximation.

For the OU and DW models we used 1000 samples from the variational approximation produced by VI to obtain the MAP estimate. We also ran the PMMH algorithm for 100,000 iterations, with a burnin of 50,000 iterations, and thinned to obtain 1000 samples representing the gold standard estimate. Here we used 300 particles. We then used these samples to obtain a corresponding gold standard MAP estimate as a baseline to compare the variational methods. We ran both VI and pathVI for 2000 iterations with $L = 50$.

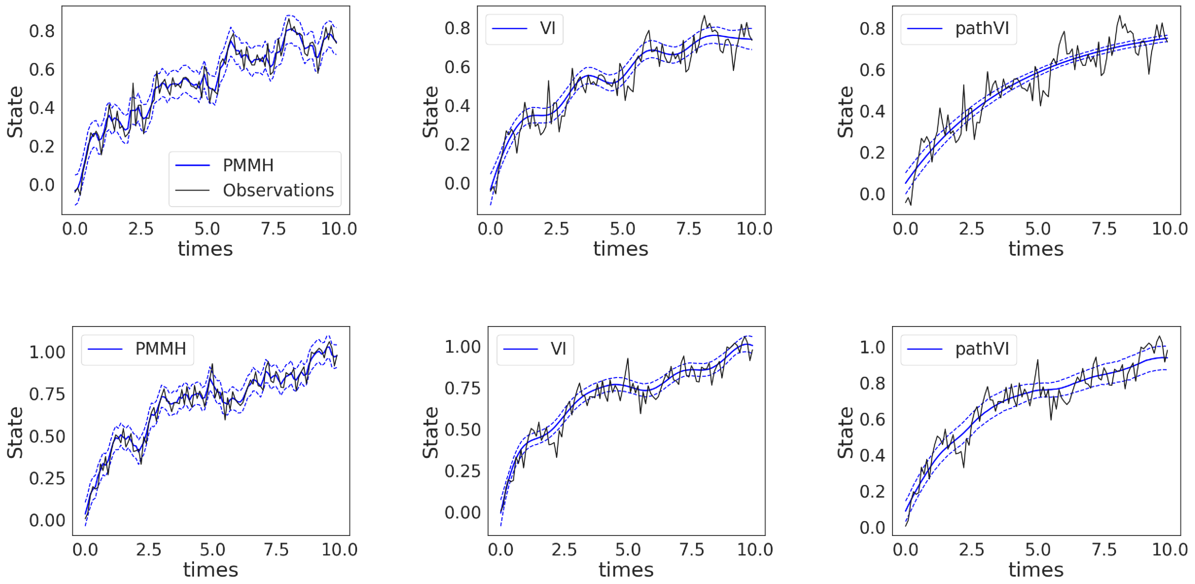


Figure 9: Summaries of the posterior distributions of the latent path of the **OU** process for two different datasets.

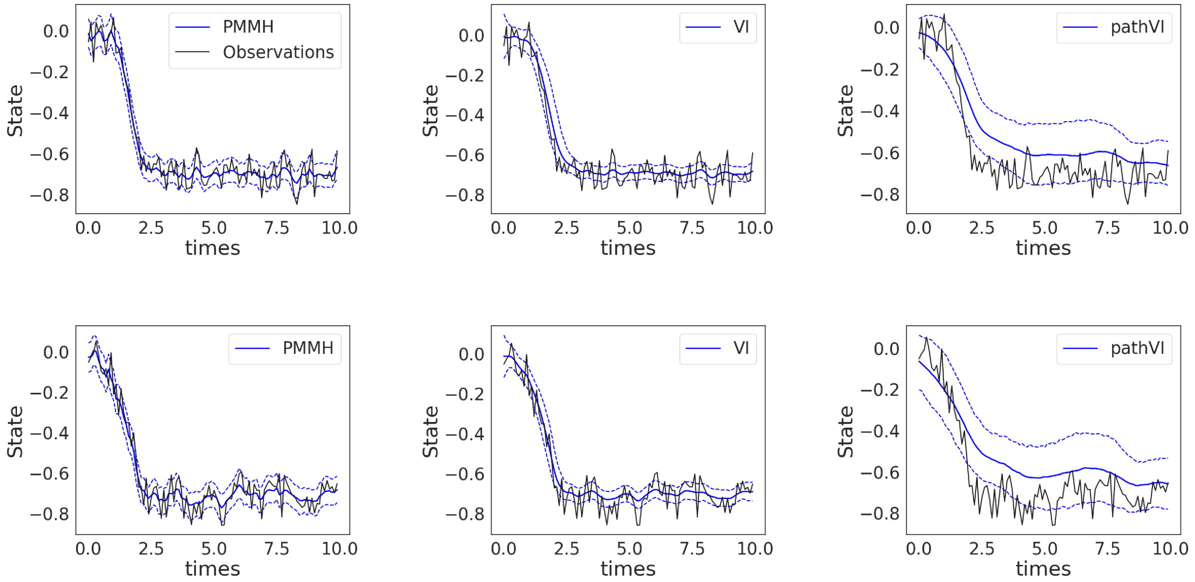


Figure 10: Summaries of the posterior distributions of the latent path of the **double-well** model for two different datasets.