Stat Comput (2015) 25:427–437
DOI 10.1007/s11222-013-9441-1

Posterior inference on parameters of stochastic differential equations via non-linear Gaussian filtering and adaptive MCMC

Simo Särkkä · Jouni Hartikainen · Isambi Sailon Mbalawata · Heikki Haario

Received: 13 November 2012 / Accepted: 26 November 2013 / Published online: 13 December 2013

Abstract This article is concerned with Bayesian estimation of parameters in non-linear multivariate stochastic differential equation (SDE) models occurring, for example, in physics, engineering, and financial applications. In particular, we study the use of adaptive Markov chain Monte Carlo (AMCMC) based numerical integration methods with non-linear Kalman-type approximate Gaussian filters for parameter estimation in non-linear SDEs. We study the accuracy and computational efficiency of gradient-free sigma-point approximations (Gaussian quadratures) in the context of parameter estimation, and compare them with Taylor series and particle MCMC approximations. The results indicate that the sigma-point based Gaussian approximations lead to better approximations of the parameter posterior distribution than the Taylor series, and the accuracy of the approximations is comparable to that of the computationally significantly heavier particle MCMC approximations.

Keywords Stochastic differential equation · Parameter estimation · Gaussian approximation · Non-linear Kalman filter · Adaptive Markov chain Monte Carlo

1 Introduction

This article is concerned with the use of non-linear Kalman-type approximate Bayesian filters in adaptive Markov chain Monte Carlo (AMCMC) estimation of the unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ in Itô stochastic differential equations (SDEs, see, e.g., Øksendal 2003) of the form

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t; \boldsymbol{\theta})\mathrm{d}t + \mathbf{L}(\mathbf{x}, t; \boldsymbol{\theta})\mathrm{d}\mathbf{W}, \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$, and $\mathbf{W}(t)$ is an s-dimensional vector of independent standard Wiener processes. Above, $\mathbf{f}(\cdot)$ is the non-linear drift function and $\mathbf{L}(\cdot)$ is the dispersion matrix of the SDE. The assumed model for the measurements $\mathbf{y}_k \in \mathbb{R}^m$ is linear Gaussian:

$$\mathbf{y}_k = \mathbf{H}(\boldsymbol{\theta})\mathbf{x}(t_k) + \mathbf{r}_k, \quad \mathbf{r}_k \sim \mathrm{N}\big(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta})\big), \tag{2}$$

where the matrices $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{R}(\boldsymbol{\theta})$ are deterministic functions of the parameters and the measurement are sampled at given time steps $t_1, \ldots, t_M$. The generalization to non-linear Gaussian measurement models is straightforward.

The central difficulty in the parameter estimation problem is that the transition density of the SDE in Eq. (1) cannot be evaluated in closed form. There exists a wide range of methods for estimating parameters of SDE models, which circumvent this problem. These include simulated maximum likelihood based methods (Pedersen 1995; Brandt and Santa-Clara 2002; Hurn et al. 2003), MCMC based methods (Jones 1998; Elerian et al. 2001; Eraker 2001; Golightly and Wilkinson 2006, 2008; Stramer et al. 2010; Stramer and Bognar 2011) as well as Exact Algorithm based methods (Beskos et al. 2006, 2009). It is also possible to directly approximate the solutions of the Fokker–Planck–Kolmogorov (FPK) equation by using standard numerical methods for solving PDEs (Hurn and Lindsay 1999; Jensen and Poulsen 2002; Jeisman 2005) such as Hermite expansions (Aït-Sahalia 2002, 2008). An overview of other methods can be found in Sørensen (2004).

Electronic supplementary material The online version of this article (doi:10.1007/s11222-013-9441-1) contains supplementary material, which is available to authorized users.

S. Särkkä (✉) · J. Hartikainen · I.S. Mbalawata · H. Haario
P.O. Box 12200, 00076 Aalto, Finland
e-mail: simo.sarkka@aalto.fi

In this article, we study another class of methods, which is based on forming a Gaussian (process) approximation to the parameter conditioned diffusion process. This approach is connected to the Taylor series approximations used in the extended Kalman filter (EKF, Jazwinski 1970) which is a widely used approach in guidance, control, target tracking, and other applications (see, e.g., Bar-Shalom et al. 2001). Given the Gaussian approximation, it is possible to evaluate the corresponding marginal likelihood of the parameters and further the (approximate) unnormalized marginal posterior density by using so-called prediction error decomposition (Schweppe 1965). Similar approximations have also been proposed in the context of linear noise approximations (LNA) of jump Markov processes and the related master equations (Kurtz 1970, 1971; Van Kampen 2007; Ferm et al. 2008; Komorowski et al. 2009; Ross et al. 2009).

Although the idea of using Gaussian approximations of diffusion processes is old (see, e.g., Kushner 1967; Jazwinski 1970), in recent years, new filtering and smoothing algorithms based on Gaussian quadrature integration and sigma-point methods have been developed (Singer 2008b; Särkkä 2006, 2007, 2010; Arasaratnam et al. 2010; Archambeau and Opper 2011; Särkkä and Sarmavuori 2013). These new methods are more accurate than the classical Taylor series based EKF methods.

Using Gaussian approximations in the context of maximum likelihood estimation of parameters in SDEs was studied by Singer (2002), who compared simulated maximum likelihood, Itô–Taylor series, and Taylor series based EKF approximations in maximum likelihood (ML) estimation of parameters. Singer (2011) extended the results and proposed the use of numerical integration and quadrature type of approximations instead of the Taylor series based EKF approximation. The use of Markov chain Monte Carlo (MCMC) methods instead of the ML estimates in the EKF based SDE parameter estimation framework was recently investigated by Mbalawata et al. (2013). To improve the performance of the basic Metropolis–Hastings based MCMC sampling, Mbalawata et al. (2013) proposed the use of Hamiltonian Monte Carlo (HMC) method. Maximum likelihood estimation in SDE models using Gauss–Hermite quadrature approximations of diffusions was also recently studied by Hurn et al. (2013).

In this article, the aim is to study the accuracy and computational requirements of Gaussian approximation based parameter estimation methods when they are combined with MCMC methods, and in particular, with adaptive MCMC methods. More specifically, we study the use of the recently developed Gaussian quadrature (sigma-point) based Gaussian approximations developed for continuous-discrete filtering and smoothing in the SDE parameter estimation problem. We evaluate the effect of the Gaussian state approximations on the parameter posterior distributions by comparing them to the grid based "exact" solutions, previously proposed Taylor series (or LNA) based approximations, and to particle MCMC approximations (Andrieu et al. 2010; Golightly and Wilkinson 2011).

2 Adaptive Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) methods (see, e.g., Liu 2001) are powerful methods for approximate computation of expectation integrals of the form

$$\mathrm{E}\big[\mathbf{g}(\boldsymbol{\theta})|\mathcal{Y}_M\big] = \int_{\mathbb{R}^d} \mathbf{g}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{Y}_M)\mathrm{d}\boldsymbol{\theta}, \tag{3}$$

where the posterior distribution $p(\boldsymbol{\theta}|\mathcal{Y}_M) \propto p(\mathcal{Y}_M|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is known only up to a normalization constant. This kind of integrals typically arises in the context of Bayesian computations (Gelman et al. 2004; Bernardo and Smith 1994). Above, $p(\mathcal{Y}_M|\boldsymbol{\theta})$ is the marginal likelihood of the measurements $\mathcal{Y}_M = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ and $p(\boldsymbol{\theta})$ is the prior distribution of the parameters $\boldsymbol{\theta} \in \mathbb{R}^d$.

The sampling properties of MCMC heavily depend on the choice of the proposal distribution and unfortunately, choosing a suitable one is hard (see, e.g., Liu 2001). In the Metropolis algorithm, a Gaussian distribution is often used, in which case the covariance matrix of the Gaussian proposal distribution needs to be well tuned. If the covariance is too small or too large, the Markov chains in MCMC will be highly correlated and hence the estimators will have a large variance (Andrieu and Thoms 2008). To overcome this problem we can use *adaptive Markov chain Monte Carlo (AMCMC)* methods where the covariance of the Gaussian proposal in the Metropolis algorithm is automatically adapted during the MCMC run (Haario et al. 1999, 2001, 2006; Liang et al. 2010; Andrieu and Thoms 2008; Roberts and Rosenthal 2007, 2009; Fort et al. 2011; Andrieu and Moulines 2006; Atchade and Rosenthal 2005; Vihola 2012).

In this article, we use the robust adaptive Metropolis (RAM) algorithm introduced by Vihola (2012), where the adaptation method is an extension of the algorithm of Haario et al. (2001). The RAM algorithm of Vihola (2012) is similar to the adaptive Metropolis algorithm of Haario et al. (2001) except that the adaptation of the covariance $\boldsymbol{\Sigma}_i$ is done in a slightly different way. RAM also has an additional mechanism to reach a given target acceptance rate, for example, the rate $\bar{\alpha}_* = 0.234$, which can be shown to be optimal under certain (ideal) conditions (Roberts and Rosenthal 2001). The RAM algorithm can be found in the supplementary material of this article.

3 Posterior inference via Bayesian filtering and smoothing

To estimate the parameters with MCMC methods, we now want to find a way to evaluate the marginal posterior probability density of the parameters $p(\boldsymbol{\theta}|\mathcal{Y}_M) \propto p(\mathcal{Y}_M|\boldsymbol{\theta})p(\boldsymbol{\theta})$ up to a normalization constant. The prior is usually easy to evaluate, but in the case of SDE parameter estimation the difficult part is the evaluation of the marginal likelihood $p(\mathcal{Y}_M|\boldsymbol{\theta})$, where the states have been integrated out. One generally applicable approach is to approximate this likelihood by using particle filtering, which leads to so-called particle Markov chain Monte Carlo (PMCMC) methods (Andrieu et al. 2010). However, their disadvantage is that they tend to require a high amount of computational resources as we will see later in this article.

The marginal likelihood $p(\mathcal{Y}_M|\boldsymbol{\theta})$ can be, in principle, computed efficiently as follows. The classical filtering theory (Jazwinski 1970) states that we can compute the prediction densities $p(\mathbf{x}(t_k)|\mathcal{Y}_{k-1}, \boldsymbol{\theta})$ and the filtering densities $p(\mathbf{x}(t_k)|\mathcal{Y}_k, \boldsymbol{\theta})$ for $k = 1, \ldots, T$ via the following recursive algorithm:

- Start from the prior $p(\mathbf{x}(t_0)|\boldsymbol{\theta}) \equiv p(\mathbf{x}(t_0)|\mathcal{Y}_0, \boldsymbol{\theta})$.
- For each measurement $\mathbf{y}_k$ for $k = 1, \ldots, M$ do:
  1. *Prediction step*: Integrate the FPK equation from the initial condition $p(\mathbf{x}(t_{k-1})|\mathcal{Y}_{k-1}, \boldsymbol{\theta})$ to time t_k, which results in the predicted density $p(\mathbf{x}(t_k)|\mathcal{Y}_{k-1}, \boldsymbol{\theta})$.
  2. *Update step*: Use Bayes' rule for computing the filtering density for step k:

$$p\big(\mathbf{x}(t_k)|\mathcal{Y}_k, \boldsymbol{\theta}\big) = \frac{p(\mathbf{y}_k|\mathbf{x}(t_k), \boldsymbol{\theta})\, p(\mathbf{x}(t_k)|\mathcal{Y}_{k-1}, \boldsymbol{\theta})}{\int p(\mathbf{y}_k|\mathbf{x}(t_k), \boldsymbol{\theta})\, p(\mathbf{x}(t_k)|\mathcal{Y}_{k-1}, \boldsymbol{\theta}) \mathrm{d}\mathbf{x}(t_k)}. \tag{4}$$

The filtering results can be used for constructing an expression for the marginal likelihood as follows. The marginal likelihood can be factored as:

$$p(\mathcal{Y}_M|\boldsymbol{\theta}) = \prod_{k=1}^{M} p(\mathbf{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}). \tag{5}$$

Given the filtering solution, the terms in the product can be computed recursively as

$$p(\mathbf{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}) = \int p\big(\mathbf{y}_k|\mathbf{x}(t_k), \boldsymbol{\theta}\big)\, p\big(\mathbf{x}(t_k)|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\big) \mathrm{d}\mathbf{x}(t_k). \tag{6}$$

Once we have obtained the expression for the marginal likelihood, we can easily evaluate the posterior distribution up to the normalization constant, which is all we need for the implementation of MCMC sampling.

Sometimes we also want to generate samples from the posterior distribution of the states. Due to the Markov properties of the state, the conditional distribution of the state given all the measurements obeys the following backward-simulation (smoothing) recursion (cf. Godsill et al. 2004):

$$p\big(\mathbf{x}(t_k)|\mathbf{x}(t_{k+1}), \mathcal{Y}_M, \boldsymbol{\theta}\big) = \frac{p(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k), \boldsymbol{\theta})\, p(\mathbf{x}(t_k)|\mathcal{Y}_k, \boldsymbol{\theta})}{p(\mathbf{x}(t_{k+1})|\mathcal{Y}_k, \boldsymbol{\theta})}, \tag{7}$$

where the right hand side only involves results from filtering. Thus once we have computed the filtering distributions, we can simulate posterior trajectories $\mathbf{x}(t_0), \ldots, \mathbf{x}(t_M)$ using the above recursion. Although we have formulated the above recursion only for the states at the measurement times, we can easily augment additional times to the trajectory and obtain samples from them as well.

Multiplying the above equation with $p(\mathbf{x}(t_{k+1})|\mathcal{Y}_M, \boldsymbol{\theta})$ and integrating over $\mathbf{x}(t_{k+1})$ gives the following backward Bayesian optimal smoothing recursion (Kitagawa 1987):

$$\begin{aligned} p\big(\mathbf{x}(t_k)|\mathcal{Y}_M, \boldsymbol{\theta}\big) &= p\big(\mathbf{x}(t_k)|\mathcal{Y}_k, \boldsymbol{\theta}\big) \\ &\quad \times \int \frac{p(\mathbf{x}(t_{k+1})|\mathbf{x}(t_k), \boldsymbol{\theta})\, p(\mathbf{x}(t_{k+1})|\mathcal{Y}_M, \boldsymbol{\theta})}{p(\mathbf{x}(t_{k+1})|\mathcal{Y}_k, \boldsymbol{\theta})} \mathrm{d}\mathbf{x}(t_{k+1}), \end{aligned} \tag{8}$$

which can be used for computing the posterior marginals of $\mathbf{x}(t_k)$ (i.e., the smoothing distributions) for $k = M - 1, \ldots, 0$.

4 Gaussian filtering and smoothing based approximation

In this section, we show how to form Gaussian approximations to the general recursion equations given in the previous section using the recent numerical approximation methods developed for state estimation (e.g., Särkkä and Solin 2012; Särkkä and Sarmavuori 2013). We derive the approximations directly using results from Itô calculus which makes the underlying approximations more explicit than the typical Euler–Maruyama discretization based derivation (e.g., Singer 2011). We also derive equations for sampling trajectories from the approximate full smoothing posterior.

Assume that after the step $k - 1$ at time t_{k-1} we have a Gaussian approximation to the filtering distribution as follows:

$$p\big(\mathbf{x}(t_{k-1})|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\big) \approx \mathrm{N}\big(\mathbf{x}(t_{k-1})|\mathbf{m}(t_{k-1}), \mathbf{P}(t_{k-1})\big), \tag{9}$$

with some mean $\mathbf{m}(t_{k-1})$ and covariance $\mathbf{P}(t_{k-1})$. We are now interested in what happens to this distribution in the

evolution from time t_{k-1} to time t_k. Because we are interested in forming a Gaussian approximation, we are specifically interested in knowing the mean and covariance of the resulting distribution. Note that the mean and covariance are actually functions of the parameter $\boldsymbol{\theta}$ as well, but we have dropped that dependence for notational convenience.

By using the generator of the SDE (e.g., Øksendal 2003) we get the following identity of an arbitrary function g:

$$\begin{aligned}&\mathrm{E}\left[\frac{\mathrm{d}g}{\mathrm{d}t}|\mathcal{Y}_{k-1},\boldsymbol{\theta}\right]\\&=\mathrm{E}\left[\sum_j\frac{\partial g}{\partial x_j}f_j(x,t)|\mathcal{Y}_{k-1},\boldsymbol{\theta}\right]\\&\quad+\frac{1}{2}\mathrm{E}\left[\sum_{ij}\left[\mathbf{L}(\mathbf{x},t;\boldsymbol{\theta})\mathbf{L}^T(\mathbf{x},t;\boldsymbol{\theta})\right]_{ij}\frac{\partial^2 g}{\partial x_i\partial x_j}|\mathcal{Y}_{k-1},\boldsymbol{\theta}\right].\end{aligned}\tag{10}$$

We are interested in the mean and covariance of the solution and so we first set $g(\mathbf{x})=x_i$, giving rise to a solution m_i, and then set $g(\mathbf{x})=(x_i-m_i)(x_j-m_j)$, which leads to the following ordinary differential equations (ODEs) for the conditional mean $\mathbf{m}(t)=\mathrm{E}[\mathbf{x}(t)|\mathcal{Y}_{k-1},\boldsymbol{\theta}]$ and covariance $\mathbf{P}(t)=\mathrm{E}[(\mathbf{x}(t)-\mathbf{m}(t))(\mathbf{x}(t)-\mathbf{m}(t))^T|\mathcal{Y}_{k-1},\boldsymbol{\theta}]$:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t}&=\mathrm{E}\big[\mathbf{f}(\mathbf{x},t;\boldsymbol{\theta})|\mathcal{Y}_{k-1},\boldsymbol{\theta}\big]\\\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t}&=\mathrm{E}\big[\big(\mathbf{x}-\mathbf{m}(t)\big)\mathbf{f}^T(\mathbf{x},t;\boldsymbol{\theta})|\mathcal{Y}_{k-1},\boldsymbol{\theta}\big]\\&\quad+\mathrm{E}\big[\mathbf{f}(\mathbf{x},t;\boldsymbol{\theta})\big(\mathbf{x}-\mathbf{m}(t)\big)^T|\mathcal{Y}_{k-1},\boldsymbol{\theta}\big]\\&\quad+\mathrm{E}\big[\mathbf{L}(\mathbf{x},t;\boldsymbol{\theta})\mathbf{L}^T(\mathbf{x},t;\boldsymbol{\theta})|\mathcal{Y}_{k-1},\boldsymbol{\theta}\big].\end{aligned}\tag{11}$$

Note that the expectations above are taken with respect to the conditional distribution of the state $p(\mathbf{x}(t)|\mathcal{Y}_{k-1},\boldsymbol{\theta})$. Thus the equations are not true ODEs in the sense that they cannot be solved without first solving the full FPK partial differential equation (Jazwinski 1970).

We now employ a simple but widely used approximation of $\mathbf{x}(t)$ as a Gaussian process such that at every time step $t\in[t_{k-1},t_k)$ we have:

$$p\big(\mathbf{x}(t)|\mathcal{Y}_{k-1},\boldsymbol{\theta}\big)\approx\mathrm{N}\big(\mathbf{x}(t)|\mathbf{m}(t),\mathbf{P}(t)\big),\tag{12}$$

where $\mathbf{m}(t)$ and $\mathbf{P}(t)$ are the mean and covariance of the process $\mathbf{x}(t)$, respectively. That is, we replace the expected values with respect to the true distribution of $\mathbf{x}(t)$ with expectations over the Gaussian approximation. The approximations to the mean and covariance integrals can then be evaluated numerically. One way is to use Taylor expansions as in extended Kalman filter (EKF) (Jazwinski 1970):

$$\begin{aligned}\mathbf{f}(\mathbf{x},t;\boldsymbol{\theta})&\approx\mathbf{f}(\mathbf{m},t;\boldsymbol{\theta})+\mathbf{F}_x(\mathbf{m},t;\boldsymbol{\theta})(\mathbf{x}-\mathbf{m})\\\mathbf{L}(\mathbf{x},t;\boldsymbol{\theta})&\approx\mathbf{L}(\mathbf{m},t;\boldsymbol{\theta}),\end{aligned}\tag{13}$$

where $\mathbf{F}_x$ is the Jacobian matrix of the mapping $\mathbf{x}\mapsto\mathbf{f}(\mathbf{x},t;\boldsymbol{\theta})$, which gives the Taylor series or LNA approximations (e.g., Mbalawata et al. 2013; Singer 2002; Ferm et al. 2008):

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}&=\mathbf{f}(\mathbf{m},t;\boldsymbol{\theta})\\\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t}&=\mathbf{P}\mathbf{F}_x^T(\mathbf{m},t;\boldsymbol{\theta})+\mathbf{F}_x(\mathbf{m},t;\boldsymbol{\theta})\mathbf{P}\\&\quad+\mathbf{L}(\mathbf{m},t;\boldsymbol{\theta})\mathbf{L}^T(\mathbf{m},t;\boldsymbol{\theta}).\end{aligned}$$

One particularly useful class of methods are Gaussian quadrature methods, also called sigma-point methods, which approximate the Gaussian integrals in the following way:

$$\int\mathbf{g}(\mathbf{x})\mathrm{N}(\mathbf{x}|\mathbf{m},\mathbf{P})\mathrm{d}\mathbf{x}\approx\sum_{i=1}^{N_\xi}W^{(i)}\mathbf{g}\big(\mathbf{m}+\sqrt{\mathbf{P}}\xi^{(i)}\big),\tag{14}$$

where $\xi^{(i)}$ are a suitable set of method specific unit sigma points and $W^{(i)}$ are some deterministic weights. The matrix square root is any matrix, which satisfies $\sqrt{\mathbf{P}}\sqrt{\mathbf{P}}^T=\mathbf{P}$. With this approximation the mean and covariance equations become (Särkkä and Sarmavuori 2013):

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}&=\sum_i W^{(i)}\mathbf{f}(\mathbf{m}+\sqrt{\mathbf{P}}\xi_i,t;\boldsymbol{\theta})\\\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t}&=\sum_i W^{(i)}\mathbf{f}(\mathbf{m}+\sqrt{\mathbf{P}}\xi_i,t;\boldsymbol{\theta})\xi_i^T\sqrt{\mathbf{P}}^T\\&\quad+\sum_i W^{(i)}\sqrt{\mathbf{P}}\xi_i\mathbf{f}^T(\mathbf{m}+\sqrt{\mathbf{P}}\xi_i,t;\boldsymbol{\theta})\\&\quad+\sum_i W^{(i)}\mathbf{L}\big(\mathbf{m}+\sqrt{\mathbf{P}}\xi^{(i)},t;\boldsymbol{\theta}\big)\mathbf{L}^T\big(\mathbf{m}+\sqrt{\mathbf{P}}\xi^{(i)},t;\boldsymbol{\theta}\big).\end{aligned}$$

One useful sigma-point method is the 3rd order spherical cubature rule ("cubature" is just another term for multidimensional quadrature) used in the cubature Kalman filter (CKF) of Arasaratnam et al. (2010), which uses $N_\xi=2n$ unit sigma points defined as

$$\xi_i=\begin{cases}\sqrt{n}\mathbf{e}_i, & i=1,\ldots,n\\-\sqrt{n}\mathbf{e}_{i-n}, & i=n+1,\ldots,2n,\end{cases}\tag{15}$$

where $\mathbf{e}_i$ denotes the unit vector to the direction of coordinate axis i, and the weights are defined as $W^{(i)}=1/(2n)$ for $i=1,\ldots,2n$. The 3rd order spherical cubature rule can be seen as a special case of the unscented transform used in the unscented Kalman filter (UKF) (Julier et al. 2000; Särkkä 2007). The third order here means that it integrates correctly any linear combination of up to third order monomials (e.g., $x_1x_2x_3$, $x_1^2x_2$, x_1^3).

One may also form the rule (14) as a Cartesian product of Gauss–Hermite quadratures as is done in the Gauss–Hermite filters (Ito and Xiong 2000; Singer 2008b). This leads to methods which are able to integrate linear combinations of higher order polynomials correctly.

By integrating the mean and covariance equations from the initial conditions $\mathbf{m}(t_{k-1})$ and $\mathbf{P}(t_{k-1})$ to time t_k, we get the following Gaussian approximation to the solution of the FPK equation:

$$p\big(\mathbf{x}(t_k)|\mathcal{Y}_{k-1},\boldsymbol{\theta}\big) \approx \mathrm{N}\big(\mathbf{x}(t_k)|\mathbf{m}\big(t_k^-\big),\mathbf{P}\big(t_k^-\big)\big). \tag{16}$$

Above, the results of prediction are denoted as $\mathbf{m}(t_k^-)$, $\mathbf{P}(t_k^-)$, where the minus at superscript means "infinitesimally before the time t_k". We use this notation to distinguish these from $\mathbf{m}(t_k)$, $\mathbf{P}(t_k)$, which denote the mean and covariance of the Gaussian approximation to $p(\mathbf{x}(t_k)|\mathcal{Y}_k,\boldsymbol{\theta})$.

Because of the Gaussianity of the above distribution and the measurement model, Bayes' rule in Eq. (4) gives

$$p\big(\mathbf{x}(t_k)|\mathcal{Y}_k,\boldsymbol{\theta}\big) \approx \mathrm{N}\big(\mathbf{x}(t_k)|\mathbf{m}(t_k),\mathbf{P}(t_k)\big), \tag{17}$$

where the mean $\mathbf{m}(t_k)$ and covariance $\mathbf{P}(t_k)$ are given by the standard Kalman filter update equations (see the supplementary material).

The conditional marginal likelihood in Eq. (6) is also Gaussian and can be evaluated simply as:

$$p(\mathbf{y}_k|\mathcal{Y}_{k-1},\boldsymbol{\theta}) \approx \mathrm{N}\big(\mathbf{y}_k|\mathbf{H}(\boldsymbol{\theta})\mathbf{m}\big(t_k^-\big),\mathbf{S}_k\big), \tag{18}$$

where $\mathbf{S}_k = \mathbf{H}(\boldsymbol{\theta})\mathbf{P}(t_k^-)\mathbf{H}^T(\boldsymbol{\theta}) + \mathbf{R}(\boldsymbol{\theta})$. The prediction error decomposition (5) then gives:

$$p(\mathcal{Y}_M|\boldsymbol{\theta}) \approx \prod_{k=1}^{M} \mathrm{N}\big(\mathbf{y}_k|\mathbf{H}(\boldsymbol{\theta})\mathbf{m}\big(t_k^-\big),\mathbf{S}_k\big). \tag{19}$$

For the smoothing recursions, we can utilize the results derived by Särkkä and Sarmavuori (2013), who give the equations for the approximate means $\mathbf{m}^s(t_k)$ and covariances $\mathbf{P}^s(t_k)$ of the (marginal) smoothing distributions in Eq. (8). The equations can be found in the supplementary material. Thus we approximately have

$$p\big(\mathbf{x}(t_k)|\mathcal{Y}_M,\boldsymbol{\theta}\big) \approx \mathrm{N}\big(\mathbf{x}(t_k)|\mathbf{m}^s(t_k),\mathbf{P}^s(t_k)\big). \tag{20}$$

The backward-simulation recursion in Eq. (7) can be obtained by formally setting $\mathbf{m}(t_{k+1}) \leftarrow \mathbf{x}(t_{k+1})$ and $\mathbf{P}(t_{k+1}) \leftarrow \mathbf{0}$ in the smoothing recursions, which gives:

$$\begin{aligned} &p\big(\mathbf{x}(t_k)|\mathbf{x}(t_{k+1}),\mathcal{Y}_M,\boldsymbol{\theta}\big) \\ &\quad\approx \mathrm{N}\big(\mathbf{x}(t_k)|\mathbf{m}(t_k) + \mathbf{G}_{k+1}\big[\mathbf{x}(t_{k+1}) - \mathbf{m}\big(t_{k+1}^-\big)\big], \\ &\qquad \mathbf{P}(t_k) - \mathbf{G}_{k+1}\mathbf{P}\big(t_{k+1}^-\big)\mathbf{G}_{k+1}^T\big). \end{aligned} \tag{21}$$

5 Experimental evaluation

5.1 Ginzburg–Landau double well potential

As the first numerical example we consider a diffusion process in a double well potential, described by the Ginzburg–Landau equation, which is a popular benchmark model for comparing different approximation schemes (e.g., Singer 2002). The Ginzburg–Landau equation reads

$$\mathrm{d}x = -\big(\alpha x + \beta x^3\big)\mathrm{d}t + \sigma\,\mathrm{d}W, \tag{22}$$

where $\boldsymbol{\theta} = (\alpha, \beta, \sigma)$ are the parameters to be estimated from data. The Ginzburg–Landau model corresponds to a diffusion at the double-well potential $\Phi(x,\boldsymbol{\theta}) = \frac{\alpha}{2}x^2 + \frac{\beta}{4}x^4$, with minima at $\pm\sqrt{-\alpha/\beta}$ when α is negative. We measure the equation with a sampling period Δt from the model

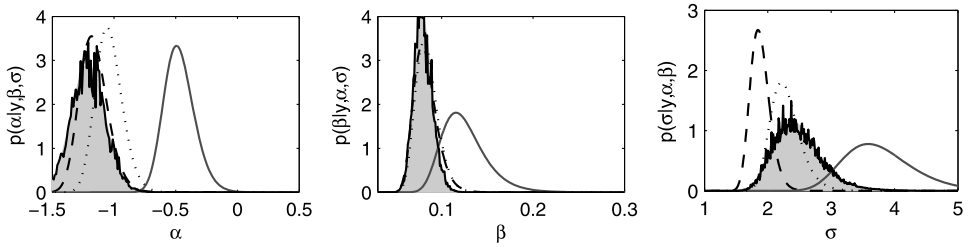
$$y_k = x(t_k) + r_k, \quad r_k \sim \mathrm{N}(0, R). \tag{23}$$

We aim to compare the performance of Taylor series, 3rd order spherical cubature rule, Gauss–Hermite quadrature in forming the Gaussian state approximation, and combine them with the adaptive MCMC method (RAM) of (Vihola 2012). As the particle MCMC method we use the particle marginal Metropolis–Hastings (PMMH) (Andrieu et al. 2010) with 5000 particles, which we also combine with the RAM adaptation method. This number of particles was chosen on the basis that the method of Doucet et al. (2012) (with the parameters fixed to true ones) suggests that 500 particles should be enough for this model—we chose still to multiply this number by ten. In the particle filter of PMMH we use the dynamic model as the importance distribution, which has the advantage that this way the evaluation of the transition density is not needed (Golightly and Wilkinson 2011). The trajectories of the SDE were simulated using the stochastic Runge–Kutta method (Kloeden and Platen 1999; Rößler 2006).

For comparison, we also implemented a benchmark filtering and smoothing solution, where the transition density is numerically solved from the FPK by using finite differences method on a relatively dense grid (the range $[-7, 7]$ with $\Delta x = 1/10$), and the Bayesian filtering and smoothing equations were approximated on the same grid via direct numerical integration. This provides a quite accurate "ground truth" solution.

We simulated a state trajectory with this model using the parameters $\boldsymbol{\theta} = (-1, 0.1, 2)$ on time interval $t \in [0, 40]$ and generated observations every $\Delta t = 2$ with variance $R = 0.1^2$. With these parameter settings the potential has minima at $\pm\sqrt{10}$, which are also the modes of the stationary (bi-modal) state distribution.

To evaluate the accuracies of different approximations in estimating the posterior distributions of the parameters we



(a) $p(\alpha|y,\beta,\sigma)$ (b) $p(\beta|y,\alpha,\sigma)$ (c) $p(\sigma|y,\alpha,\beta)$

Fig. 1 *Ginzburg–Landau: conditional posteriors of each parameter*. Here, *light gray shade* denotes the true distribution (FP-based solution), *black line* the particle MCMC estimate, *light gray line* the Taylor series-based estimate, *dotted line* Gauss–Hermite quadrature and *dashed line* spherical cubature. As can be seen, Taylor series-based method is clearly inconsistent with the true posterior, while the other methods are slightly off, but still much closer to the truth

calculated the conditional posterior distribution of each parameter in a one-dimensional grid while keeping the other parameters fixed to true values. These are plotted in Fig. 1, where we have also plotted the estimate provided by the particle MCMC method. As can be seen in the figure, the conditional distribution approximations provided by the Taylor series approach are quite far away from the true posteriors. The approximations are somewhat sensitive to particular realization of the state trajectory and measurements, but a similar phenomenon is observed each time. With the other methods the distribution approximations are much closer to the true ones although they also differ slightly from the true distribution. From the figure one can also see the stochasticity of the particle MCMC estimate, which is on average the same as the one provided by the numerical FPK solution.

We ran MCMC sampling with RAM algorithm on the unknown parameters while using the spherical cubature, Taylor series and particle MCMC based approximations. Estimates of the parameters are visualized in Fig. 2 as pair-wise marginal distributions. The FPK method was not used, because its computational requirements are far too huge to be used in MCMC sampling. As was the case with conditional distributions above, Taylor series-based method is not able to approximate the parameter posterior distribution correctly. Even though the parameter posteriors are heavily non-Gaussian, the cubature based method approximates the shape of true posterior distribution surprisingly well (assuming that PMCMC is close to the truth). The figure also shows the maximum a posteriori (MAP) estimates of parameters obtained with each of the methods separately, as well as the PMCMC-based MAP estimate, which can assumed to be close to the exact MAP estimate.

5.2 Van der Pol oscillator

As a higher dimensional non-linear example we consider parameter estimation in the Van der Pol oscillator model (see, e.g., Kandepu et al. 2008), which is described by the second order non-linear ODE

$$\frac{\mathrm{d}^2x(t)}{\mathrm{d}t^2} - \mu\big(1-\epsilon x^2(t)\big)\frac{\mathrm{d}x(t)}{\mathrm{d}t} + \omega^2 x(t) = f(t), \tag{24}$$

where the coefficients μ and ϵ as well as the angular velocity ω are unknown parameters. Above, we have also included an additional unknown forcing term $f(t)$.

We model the forcing term $f(t)$ as a sum of white noise and a stochastic resonator $c(t)$, which is formed as a sum of N harmonic components $c_n(t)$, $n = 1,\ldots,N$:

$$\frac{\mathrm{d}^2c_n(t)}{\mathrm{d}t^2} = -(n\omega_c)^2 c_n(t) + \sigma_n\epsilon_n(t), \tag{25}$$

where ω_c is the angular velocity of the force process, σ_n the strength of the noise process driving the nth harmonic, and $\epsilon_n(t)$ is a white noise process. For simplicity, we assume here that $\sigma_n = \sigma_c$ for all $n = 1,\ldots,N$. Thus, the parameter vector $\boldsymbol{\theta} = (\epsilon, \mu, \omega, \sigma, \omega_c, \sigma_c)$ is six dimensional, and the state variable $\mathbf{x} = (x, \dot{x}, c_1, \dot{c}_1, \ldots, c_N, \dot{c}_N)$ has $2+2N$ components. The full SDE model is of the form

$$\begin{aligned}
\mathrm{d}x(t) &= \dot{x}(t)\mathrm{d}t \\
\mathrm{d}\dot{x}(t) &= \mu\big(1-\epsilon x^2(t)\big)\dot{x}(t)\mathrm{d}t - \omega^2 x(t)\mathrm{d}t \\
&\quad + \big(c_1(t) + \cdots + c_N(t)\big)\mathrm{d}t \\
\mathrm{d}c_n(t) &= \dot{c}_n(t)\mathrm{d}t \\
\mathrm{d}\dot{c}_n(t) &= -(\omega_c)^2 c_n(t)\mathrm{d}t + \sigma_c\mathrm{d}W_n(t), \quad n = 1,\ldots,N.
\end{aligned} \tag{26}$$

The measurement is the state of the Van der Pol oscillator plus noise:

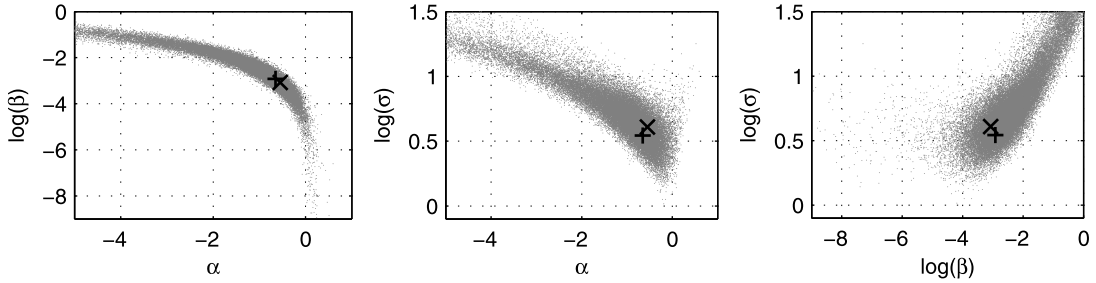
$$y_k = x(t_k) + r_k, \quad r_k \sim \mathrm{N}(0, R), \tag{27}$$

which corresponds to $\mathbf{H} = (1\ 0\ \ldots\ 0)$ in Eq. (2).

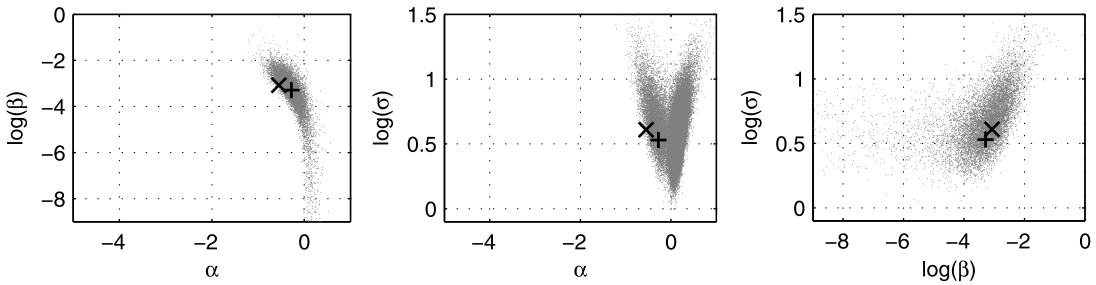
We simulated the system with parameters

$$\boldsymbol{\theta} = (1, 1/2, 1, 1/100, \pi/5, 1/100)$$

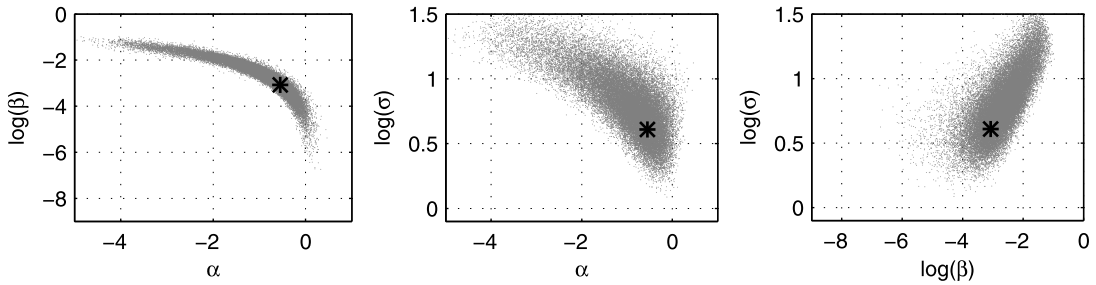
on the time interval $t \in [0, 40]$ and generated observations with sampling period $\Delta t = 1$ and variance $R = 1/10^2$. We



(a) $p(\alpha, \log(\beta)|y)$ with spherical cubature (b) $p(\alpha, \log(\sigma)|y)$ with spherical cubature (c) $p(\log(\beta), \log(\sigma)|y)$ with spherical cubature



(d) $p(\alpha, \log(\beta)|y)$ with Taylor series (e) $p(\alpha, \log(\sigma)|y)$ with Taylor series (f) $p(\log(\beta), \log(\sigma)|y)$ with Taylor series



(g) $p(\alpha, \log(\beta)|y)$ with PMCMC (h) $p(\alpha, \log(\sigma)|y)$ with PMCMC (i) $p(\log(\beta), \log(\sigma)|y)$ with PMCMC

Fig. 2 *Ginzburg–Landau: MCMC estimates of parameters*. The panels show the pair-wise marginal distributions of α, β and σ. The results for β and σ are shown in log-space since that was used in sampling. Crosses ($\times$) in the panels denote the MAP-estimates obtained from PMCMC (which should be close to the true MAP-estimates) while pluses ($+$) denote the MAP estimates of each of the methods

used $N = 2$ harmonic components in the resonator $c(t)$. Given the simulated measurements, we ran MCMC sampling with the RAM algorithm on the parameters with spherical cubature, Taylor series, and particle filter (particle MCMC) with 5000 particles approximating the marginal likelihood. This number of particles was chosen using the criterion proposed by Doucet et al. (2012). We did not test the Gauss–Hermite quadrature based method, because of its high computational requirements due to the exponential scaling of the number of computations in the state dimensionality (i.e., 7th order method would need $7^6 = 117649$ sigma-points). The FPK method is not feasible either, because it scales exponentially in the state dimensions and as stated in the previous section, it is too expensive for MCMC even in one dimension.

Figure 3 shows the pair-wise marginals of each of the parameters. It can be seen that the parameters ϵ and μ are highly correlated with each other. The parameters ω and ω_c seem to be quite well identified with all the methods, while the parameters σ and σ_c do not identify as well. The posterior approximation of the latter parameters also shows the MCMC method has difficulties in sampling of these parameters under the Gaussian approximations. However, the shapes of all the parameter posteriors are still well captured by all the methods.

5.3 Computational requirements

In this section, we evaluate the computational requirements of the methods used in the Van der Pol example with varying number of oscillator components N. The computational requirements of the different methods in parameter estimation heavily depend on the underlying algorithm used for evaluation of the marginal likelihood. They in turn depend on

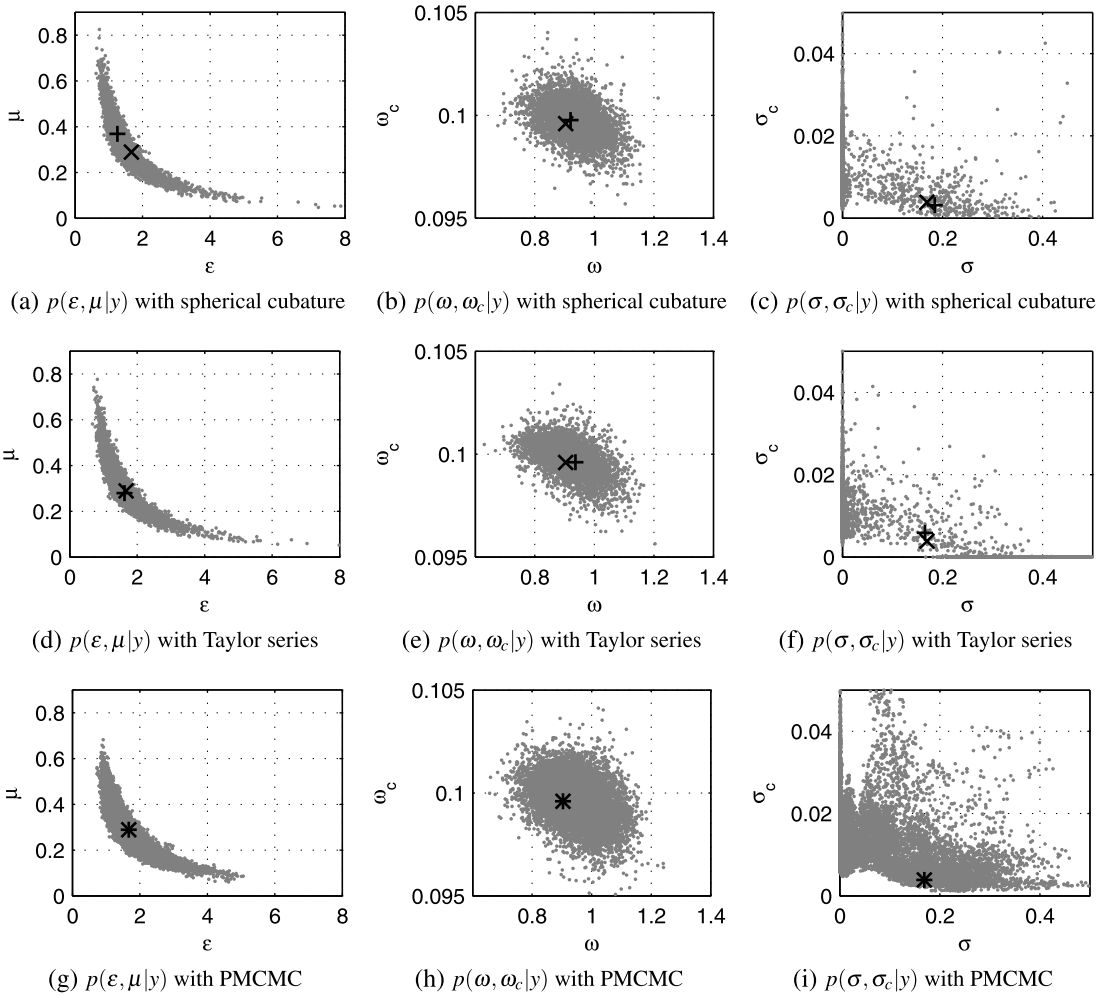


Fig. 3 *Van der Pol Oscillator: MCMC estimates of parameters.* The panels show the pair-wise marginal distributions of the parameters provided by the spherical cubature, Taylor series, and PMCMC methods. Crosses (×) in the panels denote the MAP-estimates obtained from PMCMC (assumed to be close to the true MAP-estimates) and pluses (+) denote the MAP estimates of each of the methods

the number of evaluation points (i.e., the number of sigma-points in the cubature method and particles in the particle filter), which in turn depends on the dimensionality of the state. The scaling of all the methods with the state dimensionality is roughly quadratic and they scale roughly linearly in the number of evaluation points. However, the overhead and the proportionality constants are quite different in each of the methods.

We start by investigating the number of evaluation points needed in each of the methods. We used the criterion of Doucet et al. (2012) for determining the number of particles for PMCMC (the likelihood were evaluated at the true parameter values). These and the number of evaluation points for the cubature method are shown in Table 1. The corresponding number of evaluations for the Taylor series is one although the proportionality constant is quite high. This reveals one challenge in PMCMC—the required number of particles seems to grow exponentially with the state dimension, which indeed is a well-known problem in the context of particle filtering (Snyder et al. 2008). In contrast, the number of evaluations need by the cubature method grows linearly with the number of dimensions.

Table 1 The number of evaluation points needed in the Van der Pol oscillator model as function of oscillator components N

N	PMCMC particles	Cubature sigma-points
0	500	4
1	2000	8
2	5000	12
3	50000	16
4	100000	20

The numbers of evaluation points in Table 1 do not directly give the computation time needed by the methods. To evaluate the practical speed we measured the times need for

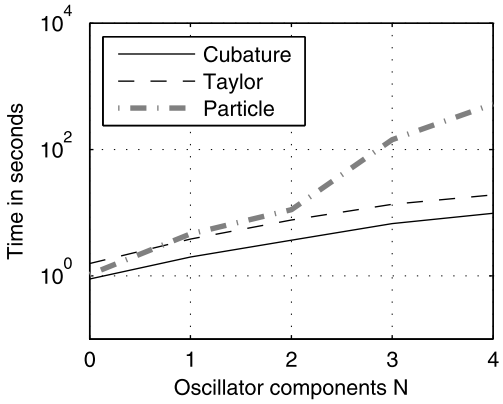


Fig. 4 Computation times as function of the number of oscillator components in the Van der Pol model. The computational requirements of the particle approximation (PMCMC) are significantly higher than those of cubature and Taylor series approximations

a single evaluation of the likelihood function with each of the methods. It is worth noting that as the evaluation time was measured using MATLAB, due to implementation details, the results slightly favor the particle filter. This would change if, for example, a C++ implementation were used.

The measured times of the likelihood evaluations are shown in Fig. 4. It can be seen that the computation time required by the particle-based method in PMCMC is much higher than the computation times of the cubature and Taylor based methods—despite the fact that the implementation favors the particle-based method. The steep scaling of the computational requirements in the particle approximation is mainly due to the almost exponential growth of the required number of particles (given in Table 1). However, without increasing the number of particles with the state dimensionality the PMCMC would not give a reasonable result at all.

6 Conclusion and discussion

In this article, we have studied the use of quadrature or sigma-point based non-linear Kalman-type Gaussian filters with adaptive Markov chain Monte Carlo (AMCMC) methods in approximate full Bayesian estimation of parameters in partially observed non-linear stochastic differential equations (SDEs). In particular, we have tested the accuracy and computational requirements of sigma-point based approximation methods against Taylor series, grid, and particle MCMC based methods. The results indicate that sigma-point based Gaussian approximations provide a surprisingly accurate approximation of the parameter posterior distribution, while the methods are computationally light-weight.

The advantage of the Gaussian approximation based methods over many other methods is that although the model family is quite general, due to the utilization of the Gaussian state approximations, the required computations remain light. The computations are order of magnitude lighter when compared to, for example, the particle MCMC method. The use of sigma-point methods makes the method black-box in the sense that to implement the method, we do not need to compute derivatives, closed form expectations or any other such quantities—being able to evaluate the drift, diffusion and measurement model functions is sufficient. Furthermore, the sigma-point approximations are more accurate than Taylor series approximations, and the use of AMCMC makes the manual tuning of proposal distributions unnecessary.

The main weakness of the methods is that they use a Gaussian approximation for the state and thus might not produce accurate results when the state posterior is strongly non-Gaussian. However, as the our experiments show, the results can be accurate even when the SDE is strongly non-linear. The non-Gaussianity of the parameter posterior is not a problem as such (cf. Fig. 2), because it is approximated using the MCMC method. The method also works well with the multivariate non-linear Van Der Pol SDE model. The results are also in good agreement with PMCMC method which produces an asymptotically exact Monte Carlo approximation of the parameter posterior, but often requires a large number of particles.

An important point to note is that the analysis of computational requirements in Sect. 5.3 is based on the implicit assumption that the accuracy of the spherical cubature rule remains the same regardless of the state dimensionality, provided that we add sigma-points with the linear rule. It is possible (and likely) that this is not strictly true. Although Gaussian approximations tend to be robust to the state dimensionality, it is possible that in some cases the exponential scaling of the space causes the integration rule to become inaccurate. However, theoretical analysis of this phenomenon is hard and heavily depends on the problem at hand.

Another thing affecting the computational requirements in practice is also the speed of mixing of the MCMC sampling. This is turn also depends on the target acceptance rate which is used in the AMCMC algorithm. For (A)PMCMC it is likely that the choice $\bar{\alpha}_* = 0.234$ which we used is not optimal. However, this choice does not affect the efficiency comparisons here.

The Gaussian approximation based methods presented in this article could be easily extended to non-linear measurement models simply by replacing the update step with the non-linear Gaussian filter update step (Ito and Xiong 2000; Wu et al. 2006; Särkkä and Sarmavuori 2013). Instead of the first order Taylor series (or LNA) it would also be possible to use higher order expansions (see, e.g., Jazwinski 1970; Maybeck 1982; Ferm et al. 2008), which might sometimes work better than the first order expansion. However, unlike the sigma-point methods, they require closed form evaluation of the derivatives of the functions and thus are not black-box in that sense. It would also be possible to use more than

the first two moments in the SDE approximation (cf. Jazwinski 1970; Maybeck 1982; Socha 2008; Singer 2008a).

7 Supplementary materials

Supplementary materials are available. Appendix A contains the full RAM algorithm, and the Kalman filter update and non-linear continuous-discrete Gaussian smoothing equations are given in Appendix B.

Acknowledgements The authors would like to thank Marko Laine for his valuable help in AMCMC methods, and the anonymous referees for their suggestions for improving the paper.

References

Aït-Sahalia, Y.: Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. Econometrica **70**(1), 223–262 (2002)

Aït-Sahalia, Y.: Closed-form expansion for multivariate diffusions. Ann. Stat. **36**(2), 906–937 (2008)

Andrieu, C., Moulines, E.: On the ergodicity properties of some adaptive MCMC algorithms. Ann. Appl. Probab. **16**(3), 1462–1505 (2006)

Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Stat. Comput. **18**(4), 343–373 (2008)

Andrieu, C., Doucet, A., Holenstein, R.: Particle Markov chain Monte Carlo methods. J. R. Stat. Soc., Ser. B, Stat. Methodol. **72**(3), 269–342 (2010)

Arasaratnam, I., Haykin, S., Hurd, T.R.: Cubature Kalman filtering for continuous-discrete systems: theory and simulations. IEEE Trans. Signal Process. **58**(10), 4977–4993 (2010)

Archambeau, C., Opper, M.: Approximate inference for continuous-time Markov processes. In: Inference and Learning in Dynamic Models. Cambridge University Press, Cambridge (2011)

Atchade, Y.F., Rosenthal, J.S.: On adaptive Markov chain Monte Carlo algorithms. Bernoulli **11**(5), 815–828 (2005)

Bar-Shalom, Y., Li, X.R., Kirubarajan, T.: Estimation with Applications to Tracking and Navigation. Wiley, New York (2001)

Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, New York (1994)

Beskos, A., Papaspiliopoulos, O., Roberts, G., Fearnhead, P.: Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). J. R. Stat. Soc., B **68**(3), 333–382 (2006)

Beskos, A., Papaspiliopoulos, O., Roberts, G.: Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. Ann. Stat. **37**(1), 223–245 (2009)

Brandt, M., Santa-Clara, P.: Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. J. Financ. Econ. **63**(2), 161–210 (2002)

Doucet, A., Pitt, M., Kohn, R.: Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator (2012). arXiv:1210.1871

Elerian, O., Chib, S., Shephard, N.: Likelihood inference for discretely observed non-linear diffusions. Econometrica **69**(4), 959–993 (2001)

Eraker, B.: MCMC analysis of diffusion models with application to finance. J. Bus. Econ. Stat. **19**(2), 177–191 (2001)

Ferm, L., Lötstedt, P., Hellander, A.: A hierarchy of approximations of the master equation scaled by a size parameter. J. Sci. Comput. **34**(2), 127–151 (2008)

Fort, G., Moulines, E., Priouret, P.: Convergence of adaptive MCMC algorithms: ergodicity and law of large number. Tech. Rep., LTCI (2011)

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, 2nd edn. CRC Press/Chapman & Hall, Boca Raton/London (2004)

Godsill, S.J., Doucet, A., West, M.: Monte Carlo smoothing for non-linear time series. J. Am. Stat. Assoc. **99**(465), 156–168 (2004)

Golightly, A., Wilkinson, D.J.: Bayesian sequential inference for nonlinear multivariate diffusions. Stat. Comput. **16**(4), 323–338 (2006)

Golightly, A., Wilkinson, D.J.: Bayesian inference for nonlinear multivariate diffusion models observed with error. Comput. Stat. Data Anal. **52**(3), 1674–1693 (2008)

Golightly, A., Wilkinson, D.J.: Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. Interface Focus **1**(6), 807–820 (2011)

Haario, H., Saksman, E., Tamminen, J.: Adaptive proposal distribution for random walk Metropolis algorithm. Comput. Stat. **14**(3), 375–395 (1999)

Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**(2), 223–242 (2001)

Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: efficient adaptive MCMC. Stat. Comput. **16**(4), 339–354 (2006)

Hurn, A.S., Lindsay, K.A.: Estimating the parameters of stochastic differential equations. Math. Comput. Simul. **48**(4–6), 373–384 (1999)

Hurn, A.S., Lindsay, K.A., Martin, V.L.: On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations. J. Time Ser. Anal. **24**(1), 45–63 (2003)

Hurn, A.S., Lindsay, K.A., McClelland, A.J.: A quasi-maximum likelihood method for estimating the parameters of multivariate diffusions. J. Econom. **172**, 106–126 (2013)

Ito, K., Xiong, K.: Gaussian filters for nonlinear filtering problems. IEEE Trans. Autom. Control **45**(5), 910–927 (2000)

Jazwinski, A.H.: Stochastic Processes and Filtering Theory. Academic Press, San Diego (1970)

Jeisman, J.: Estimation of the Parameters of Stochastic Differential Equations. Doctoral dissertation, Queensland University of Technology (2005)

Jensen, B., Poulsen, R.: Transition densities of diffusion processes: numerical comparison of approximation techniques. J. Deriv. **9**(4), 18–32 (2002)

Jones, C.S.: A simple Bayesian method for the analysis of diffusion processes. Working paper, University of Pennsylvania (1998)

Julier, S.J., Uhlmann, J.K., Durrant-Whyte, H.F.: A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Trans. Autom. Control **45**(3), 477–482 (2000)

Kandepu, R., Foss, B., Imsland, L.: Applying the unscented Kalman filter for nonlinear state estimation. J. Process Control **18**(7–8), 753–768 (2008)

Kitagawa, G.: Non-Gaussian state-space modeling of nonstationary time series. J. Am. Stat. Assoc. **82**(400), 1032–1041 (1987)

Kloeden, P.E., Platen, E.: Numerical Solution to Stochastic Differential Equations. Springer, Berlin (1999)

Komorowski, M., Finkenstädt, B., Harper, C., Rand, D.: Bayesian inference of biochemical kinetic parameters using the linear noise approximation. BMC Bioinform. **10**(1), 343 (2009)

Kurtz, T.G.: Solutions of ordinary differential equations as limits of pure jump Markov processes. J. Appl. Probab. **7**(1), 49–58 (1970)

Kurtz, T.G.: Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. J. Appl. Probab. **8**(2), 344–356 (1971)

Kushner, H.J.: Approximations to optimal nonlinear filters. IEEE Trans. Autom. Control **12**(5), 546–556 (1967)

Liang, F., Liu, C., Carroll, R.J.: Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples. Wiley, Chichester (2010)

Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, Berlin (2001)

Maybeck, P.: Stochastic Models, Estimation and Control vol. 2. Academic Press, San Diego (1982)

Mbalawata, I.S., Särkkä, S., Haario, H.: Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering. Comput. Stat. **28**(3), 1195–1223 (2013)

Øksendal, B.: Stochastic Differential Equations: An Introduction with Applications. Springer, Berlin (2003)

Pedersen, A.R.: A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scand. J. Stat. **22**(1), 55–71 (1995)

Roberts, G.O., Rosenthal, J.S.: Optimal scaling for various Metropolis–Hastings algorithms. Stat. Sci. **16**(4), 351–367 (2001)

Roberts, G.O., Rosenthal, J.S.: Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. Applied probability **44**(2), 458–475 (2007)

Roberts, G.O., Rosenthal, J.S.: Examples of adaptive MCMC. J. Comput. Graph. Stat. **18**(2), 349–367 (2009)

Ross, J.V., Pagendam, D.E., Pollett, P.K.: On parameter estimation in population models II: multi-dimensional processes and transient dynamics. Theor. Popul. Biol. **75**(2), 123–132 (2009)

Rößler, A.: Runge-Kutta methods for Ito stochastic differential equations with scalar noise. BIT Numer. Math. **46**, 97–110 (2006)

Särkkä, S.: Recursive Bayesian inference on stochastic differential equations. Doctoral dissertation, Helsinki University of Technology (2006)

Särkkä, S.: On unscented Kalman filtering for state estimation of continuous-time nonlinear systems. IEEE Trans. Autom. Control **52**(9), 1631–1641 (2007)

Särkkä, S.: Continuous-time and continuous-discrete-time unscented Rauch-Tung-Striebel smoothers. Signal Process. **90**(1), 225–235 (2010)

Särkkä, S., Sarmavuori, J.: Gaussian filtering and smoothing for continuous-discrete dynamic systems. Signal Process. **93**, 500–510 (2013)

Särkkä, S., Solin, A.: On continuous-discrete cubature Kalman filtering. In: Proceedings of SYSID 2012, pp. 1210–1215 (2012)

Schweppe, F.C.: Evaluation of likelihood functions for Gaussian signals. IEEE Trans. Inf. Theory **11**, 61–70 (1965)

Singer, H.: Parameter estimation of nonlinear stochastic differential equations: simulated maximum likelihood versus extended Kalman filter and Itô-Taylor expansion. J. Comput. Graph. Stat. **11**, 972–995 (2002)

Singer, H.: Generalized Gauss-Hermite filtering. AStA Adv. Stat. Anal. **92**(2), 179–195 (2008a)

Singer, H.: Nonlinear continuous time modeling approaches in panel research. Stat. Neerl. **62**(1), 29–57 (2008b)

Singer, H.: Continuous-discrete state-space modeling of panel data with nonlinear filter algorithms. AStA Adv. Stat. Anal. **95**(4), 375–413 (2011)

Snyder, C., Bengtsson, T., Bickel, P., Anderson, J.: Obstacles to high-dimensional particle filtering. Mon. Weather Rev. **136**(12), 4629–4640 (2008)

Sørensen, H.: Parametric inference for diffusion processes observed at discrete points in time: a survey. Int. Stat. Rev. **72**(3), 337–354 (2004)

Socha, L.: Linearization Methods for Stochastic Dynamic Systems. Springer, Berlin (2008)

Stramer, O., Bognar, M.: Bayesian inference for irreducible diffusion processes using the pseudo-marginal approach. Bayesian Anal. **6**(2), 231–258 (2011)

Stramer, O., Bognar, M., Schneider, P.: Bayesian inference for discretely sampled Markov processes with closed-form likelihood expansions. J. Financ. Econ. **8**(4), 450–480 (2010)

Van Kampen, N.G.: Stochastic processes in physics and chemistry, 3rd edn. Elsevier, Amsterdam (2007)

Vihola, M.: Robust adaptive Metropolis algorithm with coerced acceptance rate. Stat. Comput. **22**(5), 997–1008 (2012)

Wu, Y., Hu, D., Wu, M., Hu, X.: A numerical-integration perspective on Gaussian filters. IEEE Trans. Signal Process. **54**(8), 2910–2921 (2006)