Statistical Inference for Stochastic Processes (2021) 24:353–403
https://doi.org/10.1007/s11203-021-09239-3

A Kalman particle filter for online parameter estimation with applications to affine models

Jian He[1,2] · Asma Khedher[1] · Peter Spreij[1,3]

Received: 17 February 2020 / Accepted: 11 February 2021 / Published online: 3 March 2021

Abstract
In this paper we address the problem of estimating the posterior distribution of the static parameters of a continuous-time state space model with discrete-time observations by an algorithm that combines the Kalman filter and a particle filter. The proposed algorithm is semi-recursive and has a two layer structure, in which the outer layer provides the estimation of the posterior distribution of the unknown parameters and the inner layer provides the estimation of the posterior distribution of the state variables. This algorithm has a similar structure as the so-called recursive nested particle filter, but unlike the latter filter, in which both layers use a particle filter, our algorithm introduces a dynamic kernel to sample the parameter particles in the outer layer to obtain a higher convergence speed. Moreover, this algorithm also implements the Kalman filter in the inner layer to reduce the computational time. This algorithm can also be used to estimate the parameters that suddenly change value. We prove that, for a state space model with a certain structure, the estimated posterior distribution of the unknown parameters and the state variables converge to the actual distribution in L^p with rate of order $\mathcal{O}(N^{-\frac{1}{2}} + \Delta^{\frac{1}{2}})$, where N is the number of particles for the parameters in the outer layer and Δ is the maximum time step between two consecutive observations. We present numerical results of the implementation of this algorithm, in particularly we implement this algorithm for affine interest models, possibly with stochastic volatility, although the algorithm can be applied to a much broader class of models.

Keywords Affine process · State space model · Kalman filter · Particle filter · Parameter estimation · Posterior distribution

✉ Jian He
j.he2@uva.nl

Asma Khedher
a.khedher@uva.nl

Peter Spreij
spreij@uva.nl

[1] Korteweg-de Vries Institute, University of Amsterdam, PO Box 94248, 1090GE Amsterdam, The Netherlands

[2] ABN AMRO Bank N.V., Amsterdam, The Netherlands

[3] IMAPP, Radboud University, Nijmegen, The Netherlands

Mathematics Subject Classification 62P05 · 65C35 · 93E11

1 Introduction

We pose the problem, describe its background and give a brief sketch of earlier approaches. After that we explain our approach and contribution to the literature and outline the organization of the present paper.

1.1 Problem description and background

When using stochastic models in a practical environment, model parameters need to be estimated, which turns out to be a very challenging problem. In the present paper we present a method based on a combination of the Kalman and particle filters to do so. In Statistics one can discern two paradigms, Bayesian and frequentist, with ensuing Bayesian and frequentist estimation methods. Among the latter, Maximum Likelihood estimation (MLE) is a classical one. Such methods can also be categorized as online or offline depending on whether the data are used sequentially, or used in batches of observations. The MLE approach is to find the estimate which maximizes the likelihood function of the observed data. The Bayesian approach, however, considers the parameters as random variables. Prior distributions of these, reflecting prior knowledge of the parameters, is updated by the observations through the likelihood, resulting in posterior distributions.In applications, the MLE based on offline methods is often linked to the Kalman filter or its modifications such as the extended Kalman filter, see Einicke and White (1999) and Wan and Nelson (2002), or the unscented Kalman filter, see Wan and van der Merwe (2002), because these algorithms can compute or approximate the likelihood function analytically. However, a common problem of the MLE calibration is that the likelihood function is usually not convex. Hence the numerical optimization of the likelihood often ends up at a local maximum instead of the global maximum. This problem can be even more severe when dealing with models with many parameters, such as multi-factor Hull–White and Vasiěk models, popular in interest rate modeling. Originally, the MLE method requires static model parameters, while in reality the model parameters, such as volatility in financial models, could change over time. Later on, likelihood based methods have been developed to cope with this situation as well. Change point methods, see for example Nemeth et al. (2014), are developed to address the abrupt change of parameters, but those models require a separate model to determine when the model parameters change in the time series, which could increase the complexity of the model. An alternative is to use online methods, which have received more and more attention in recent decades.

Attempts to solve the problem of estimating the static parameters online was to include simulations (particles) of parameter values. One then has a *particle filter*, see for example Doucet et al. (2000), Gordon et al. (1993), Kitagawa (1996), Liu and Chen (1998) and Kantas et al. (2015) for a survey. However, through successive time steps this approach can quickly lead to what is called particle degeneracy of the parameter space. One solution to this degeneracy problem is to use a kernel density to estimate the posterior distribution of the parameters from which new parameter particles can be drawn at each time step (Liu and West 2001). Besides the fact that a study of the convergence of the algorithm introduced in this latter paper has not been carried out, also such a method could deliver poor performance in some simple set-ups other than the low dimensional case (Miguez et al. 2005).

In recent years, some new methods have been proposed to deal with the online parameter estimation problem, including the iterated batch importance sampling (IBIS), see Chopin (2002), the sequential Monte Carlo square (SMC2) simulation, see Chopin et al. (2013), and the recursive nested particle filter (RNP filter, also RNPF in short), see Crisan and Míguez (2018). The SMC2 and the RNPF use two layers of Monte Carlo methods to overcome certain difficulties with the IBIS method, see Papavasiliou (2006). An important difference between SMC2 and the RNPF is that the SMC2 is a non-recursive method, whereas the RNPF is recursive (see Ljung and Söderström 1983 for a definition of a recursive (online) algorithm for the estimation of a given model parameter). Hence in general, RNPF is more efficient than SMC2.

In Crisan and Míguez (2018) the estimated posterior measure of the parameters by using an RNPF algorithm is shown to converge to the actual measure in L^p-norm with rate $N^{-1/2} + M^{-1/2}$, where N is the number of particles for the parameter estimation in the outer layer and $N \times M$ is the number of particles for the state variables in the inner layer. The RNPF has some drawbacks for a practical application. One is that the computation of the two Monte Carlo layers is very time consuming, another one is that the RNPF requires that the parameter mutation size is small enough. As a consequence the RNPF converges very slowly to the actual value of the parameters and hence requires a very long time series of data, which is very often not available in many applications. To obtain a faster while still accurate algorithm, we propose an algorithm that combines the Kalman filter and a particle filter together with a so-called jittering kernel. The mixture of a Kalman filter and a particle filter is previously considered in Andrieu and Doucet (2002) and Chen and Liu (2000) for conditional linear-Gaussian systems, although the focus in these papers is different, merely on filtering with known parameters while in the present paper we focus on parameter estimation. Other algorithms have been proposed by Stroud et al. (2018) that combine an Ensemble Kalman filter for state estimation with various approximations (one involving a particle filter too) for the updates for the parameter posterior. Their performances have only been numerically evaluated in examples. We apply our algorithm to well chosen non-Gaussian non-linear models and we also provide a convergence analysis of the parameter estimators, an issue not treated in Andrieu and Doucet (2002), Chen and Liu (2000) and Stroud et al. (2018). More explanation follows in the next section.

1.2 Contribution

In this paper, we consider joint parameter and state estimation for a state space model where the state evolves *continuously* in time, whereas the observations are sampled at *discrete* time instants. We use a Bayesian online approach to parameter estimation. We propose an algorithm which combines the Kalman filter and a particle filter for online estimation of the posterior distribution of the unknown parameters. This algorithm has a similar structure as the RNPF, it is a semi-recursive algorithm with also two layer structure: the inner layer provides the approximation on the posterior distribution of the state variables conditioned on the parameter particles generated in the outer layer, while the outer layer provides an approximation of the posterior distribution of the parameters by using the outcome of the inner layer.

Our proposed methodology has two main differences when compared to the RNPF algorithm. One difference is that in the inner layer, the posterior distribution of the state variables is estimated by the Kalman filter instead of a particle filter. The implementation of the Kalman filter reduces the computational complexity and hence results in a much faster and robust algorithm. The second difference is in the outer layer. In the RNPF the parameter samples are generated from a certain kernel function. In order to obtain a recursive algorithm, some requirements on the kernel function are introduced. This results in a kernel that significantly reduces the convergence speed of the RNPF. We overcome this problem by using dynamic jittering kernels. Especially in this paper, we implement two different kernel functions. One is applied at the beginning stage to obtain a higher convergence speed. The consequence, however, is that the algorithm is not recursive at this beginning stage since this kernel function does not satisfy the requirements of a recursive algorithm. The other kernel is applied when the variance of the parameter particles decreases to a certain level which is such that this kernel function satisfies the conditions for a recursive method. From that time on, the algorithm is truly recursive. From the numerical experiments we performed, we observe that the variance of the particles decreases very fast at the beginning stage, usually after hundreds steps. Hence by using these two different kernel functions, the algorithm converges much faster than the RNPF.

This paper also provides theoretical results on the asymptotical behavior of the proposed algorithm. When dealing with linear and Gaussian state space models, we show that our algorithm converges with a speed of order $\mathcal{O}(N^{-1/2})$, where N is the number of particles for the parameter space. When dealing with non-Gaussian or non-linear models, the Kalman filter in the inner layer could produce a biased estimate of the posterior distribution of the state variables. This makes it in general difficult to study the convergence of the posterior distribution of the parameters. Although it is shown in Pérez-Vieites et al. (2018) that, under certain assumptions, the bias introduced in the inner layer makes the posterior distribution of the parameters converge to a biased distribution, this bias is intractable in general. In this paper, for models with a well chosen structure (affine models for interest rates), we show that the estimated distributions of the parameters and the states converge to the actual distributions in L^p with rate of order $\mathcal{O}(N^{-1/2} + \Delta^{1/2})$ under certain regularity assumptions, where Δ is the maximum time step between consecutive observations. Note that we do not have to deal with particles in the inner layer, which improves on the order $M^{-1/2}$ term for convergence rate of the RNPF. Our proofs are inspired by those in Crisan and Míguez (2018), but at crucial steps we obtain novel results. These are due to the use of the Kalman filter in one of the layers and to the size of the time discretization that governs the observations of the continuous-time system, the latter not playing a role in the setting of the cited reference.

To illustrate the performance of the algorithm, we present numerical results of the parameter estimation on several affine interest rate models, some allowing for stochastic volatility, including the Vasiček model, also known as the two-factor Hull-White model with constant parameters, and the Cox-Ingersoll-Ross (CIR) model. For the CIR model we have also implemented the RNPF and we observed that our algorithm outperforms the RNPF. Although the algorithm is designed for static parameter estimation, it can also be used to estimate parameters that perform sudden changes in value. We present an implementation of the algorithm in such a situation, and we observe that the algorithm is able to quickly track such a sudden change.

1.3 Organization of the paper

In Sect. 2 we present the state space model of interest. This section also provides brief reviews on Bayesian filters, including the Kalman filter and the particle filter, and online parameter estimation using particle filters. Section 3 contains an encompassing framework for various affine models that are used in interest rate modeling and to which we apply our proposed Kalman particle algorithm, which is introduced in Sect. 4. In Sect. 5 we provide the convergence analysis and in Sect. 6 the numerical results are presented. Finally, Sect. 7 is devoted to the conclusions. In the "Appendix" we collect some background results on affine processes.

1.4 Notation

Let $d \geq 1$, $S \subseteq \mathbb{R}^d$, and $\mathcal{B}(S)$ be the sigma algebra of Borel subsets of S. We denote by $\mathbf{1}_A$ the indicator function on $A \in \mathcal{B}(S)$ and by δ_x the Dirac measure for a given $x \in S$, i.e.,

$$\delta_x(A) = \mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

Suppose given a function $f : S \to \mathbb{R}$ and a probability measure μ on $(S, \mathcal{B}(S))$. We denote the integral of f w.r.t. μ by $(f, \mu) := \int_S f(x)\, \mu(\mathrm{d}x)$ and the supremum norm of f by $\|f\|_\infty = \sup_{x \in S} |f(x)|$. In the case of a conditional measure $\nu(\cdot \mid y)$, $y \in S$, defined on $(S, \mathcal{B}(S))$, we use the notation $(f, \nu(\cdot \mid y)) := \int_S f(x)\, \nu(\mathrm{d}x \mid y)$.

We use the notation $x_{0:k} := (x_0, \ldots, x_k)$ for a discrete-time sequence up to time k of a process $(x_k)_{k \in \mathbb{N}}$. By $\cdot^\top$, we denote the transpose of a vector or a matrix. The Euclidian norm of an element $x \in \mathbb{R}^d$, is denoted by $\|x\|$ and the L^p-norm, for $p \geq 1$ of a random variable X, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is denoted by $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$. Densities of random variables or vectors x (always assumed to exist w.r.t. the Lebesgue measure) are often denoted p, or $p(x)$ and conditional densities of X given $Y = y$ are often denoted $p(x \mid y)$, possibly endowed with sub- or superscripts.

2 Set up and background on parameters estimation using filters

In this section we outline the set up, we pose the problem formulation, give a brief survey of various filters (Bayesian, Kalman, particle filter) and address the parameter estimation problem using particle filters. Time is assumed to be discrete.

2.1 Discrete-time state space model

We consider the following general *state space* model, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

$$\begin{aligned} x_k &= f_k(x_{k-1}, u_k) \quad k \in \mathbb{N}^+, \\ y_k &= h_k(x_k, v_k)\,, \quad k \in \mathbb{N}^+\,, \end{aligned} \tag{2.1}$$

where $f_k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, $h_k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^m$ are given functions and $\{u_k\}_{k \in \mathbb{N}^+}$ and $\{v_k\}_{k \in \mathbb{N}^+}$ are d-dimensional strong white noise processes, so independent sequences, both independent of the initial condition x_0, and mutually independent as well. Parameters in the functions f_k and h_k, together with the covariance of u_k and v_k can be seen as the parameters of the state space model, and to which we refer to as θ.

It follows that the model (2.1) satisfies the properties of a stochastic system, i.e. at every (present) time $k \geq 1$ the future states and future observations (x_j, y_j), $j \geq k$, are conditionally independent from the past states and observations (x_j, y_{j-1}), $j \leq k$, given the present state x_k, see van Schuppen (1989). It then follows that $\{x_k\}_{k \in \mathbb{N}}$ is a Markov process, and for every $k \geq 1$ one has that y_k and $y_{1:k-1}$ are conditionally independent given x_{k-1}, in terms of densities,

$$p(y_k \mid y_{1:k-1}, x_k) = p(y_k \mid x_k), \quad \text{for} \quad k \in \mathbb{N}^+ . \tag{2.2}$$

Moreover, one also has, for every $k \geq 1$, that x_k and $y_{1:k-1}$ are conditionally independent given x_k, in terms of densities,

$$p(x_k \mid x_{k-1}, y_{1:k-1}) = p(x_k \mid x_{k-1}). \tag{2.3}$$

The latter equation has the consequence

$$p(x_k \mid y_{1:k-1}) = \int p(x_k \mid x_{k-1}) p(x_{k-1} \mid y_{1:k-1}) \mathrm{d}x_{k-1}. \tag{2.4}$$

We are interested in estimating the (latent) state process $\{x_k\}_{k \in \mathbb{N}^+}$, but only have access to the process $\{y_k\}_{k \in \mathbb{N}^+}$ which represents the observations. Because of the existence of the white noise in the data, estimating the value of the latent states $\{x_k\}_{k \in \mathbb{N}^+}$ by the observations $\{y_k\}_{k \in \mathbb{N}^+}$ is not trivial. There are different methodologies in the literature to estimate the latent process (see e.g. Press 2003; Chui and Chen 2017; Arulampalam et al. 2002). We introduce some of these methodologies in our paper since we will need them in our analysis later. We first introduce the Bayesian filter.

2.2 Bayesian filter of discrete-time Markovian state space model

The *Bayesian filter*, see e.g. Press (2003), Robert (2007) for an overview, is used to estimate the latent states $\{x_k\}_{k \in \mathbb{N}^+}$ in (2.1) given the parameter θ. We define the initial probability measure π_0 of x_0, and the transition measure π_k^θ of x_k under a given parameter θ at time k by

$$\begin{aligned} \pi_0(A) &= \mathbb{P}(x_0 \in A), \\ \pi_k^\theta(A \mid x_{k-1}) &= \mathbb{P}(x_k \in A \mid x_{k-1}, \theta), \qquad k \in \mathbb{N}^+ , \end{aligned} \tag{2.5}$$

where $A \in \mathcal{B}(\mathbb{R}^d)$ is a Borel set.

The methodology in Bayesian filtering consists of two parts: prediction and update. At every time point k, the prediction part computes (estimates) the *prior measure* of x_k (a time k given the past observations up to time $k-1$) and the update part computes (estimates) the *posterior measure* of x_k given the past up to time k, respectively given by

$$\begin{aligned} \gamma_k^\theta(\mathrm{d}x_k) &= \mathbb{P}(\mathrm{d}x_k \mid y_{1:k-1}, \theta), \\ \Gamma_k^\theta(\mathrm{d}x_k) &= \mathbb{P}(\mathrm{d}x_k \mid y_{1:k}, \theta), \qquad k \in \mathbb{N}^+ . \end{aligned} \tag{2.6}$$

Using Bayes' rule, we deduce that the density function of the prior distribution is given by

$$\begin{aligned} p(x_k \mid y_{1:k-1}, \theta) &= \int p(x_k \mid x_{k-1}, y_{1:k-1}, \theta) p(x_{k-1} \mid y_{1:k-1}, \theta) \, \mathrm{d}x_{k-1} \\ &= \int p(x_k \mid x_{k-1}, \theta) p(x_{k-1} \mid y_{1:k-1}, \theta) \, \mathrm{d}x_{k-1} , \end{aligned}$$

where we used (2.3) to get the last equality. This implies the relation

$$\gamma_k^\theta(\mathrm{d}x_k) = \int \pi_k^\theta(\mathrm{d}x_k \mid x_{k-1}) \Gamma_{k-1}^\theta(\mathrm{d}x_{k-1}). \tag{2.7}$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an integrable function w.r.t. the measure γ_k^θ. Then we get by Fubini's theorem

$$\begin{aligned}(f, \gamma_k^\theta) &= \int\int f(x_k)\pi_k^\theta(\mathrm{d}x_k \mid x_{k-1})\Gamma_{k-1}^\theta(\mathrm{d}x_{k-1})\\ &= \int (f, \pi_k^\theta(\cdot \mid x_{k-1}))\Gamma_{k-1}^\theta(\mathrm{d}x_{k-1}),\end{aligned}$$

which we abbreviate by

$$(f, \gamma_k^\theta) = ((f, \pi_k^\theta), \Gamma_{k-1}^\theta). \tag{2.8}$$

The purpose of the Bayesian algorithm is to sequentially compute the posterior measure Γ_k^θ. Let

$$l_{y_k}^\theta(x) = p(y_k \mid x, \theta)$$

be the density (with some abuse of statistical terminology we often also call it likelihood) of the realized observation y_k conditional on the state value $x_k = x$ and the model parameter θ. Then using Bayes' rule, (2.2) and (2.7), we obtain for a function f that is integrable w.r.t. Γ_k^θ

$$\begin{aligned}(f, \Gamma_k^\theta) &= \int f(x_k)p(x_k \mid y_{1:k}, \theta)\,\mathrm{d}x_k\\ &= \int f(x_k)\frac{p(x_k, y_k, y_{1:k-1} \mid \theta)}{p(y_{1:k} \mid \theta)}\,\mathrm{d}x_k\\ &= \frac{\int f(x_k)p(y_k \mid x_k, y_{1:k-1}, \theta)p(x_k \mid y_{1:k-1}, \theta)\,\mathrm{d}x_k}{p(y_k \mid y_{1:k-1}, \theta)}\\ &= \frac{\int f(x_k)p(y_k \mid x_k, \theta)p(x_k \mid y_{1:k-1}, \theta)\,\mathrm{d}x_k}{\int p(y_k \mid x_k, \theta)p(x_k \mid y_{1:k-1}, \theta)\,\mathrm{d}x_k}\\ &= \frac{\int f(x_k)l_{y_k}^\theta(x_k)p(x_k \mid y_{1:k-1}, \theta)\,\mathrm{d}x_k}{\int l_{y_k}^\theta(x_k)p(x_k \mid y_{1:k-1}, \theta)\,\mathrm{d}x_k}\\ &= \frac{\int f(x_k)l_{y_k}^\theta(x_k)\int \pi_k^\theta(\mathrm{d}x_k \mid x_{k-1})\Gamma_{k-1}^\theta(\mathrm{d}x_{k-1})}{\int l_{y_k}^\theta(x_k)\int \pi_k^\theta(\mathrm{d}x_k \mid x_{k-1})\Gamma_{k-1}^\theta(\mathrm{d}x_{k-1})},\end{aligned}$$

which we abbreviate, similar to (2.8), by

$$(f, \Gamma_k^\theta) = \frac{((f l_{y_k}^\theta, \pi_k^\theta), \Gamma_{k-1}^\theta)}{((l_{y_k}^\theta, \pi_k^\theta), \Gamma_{k-1}^\theta)}. \tag{2.9}$$

If we assume the likelihood function $l_{y_k}^\theta$ and the transition measure π_k^θ are known, then given the posterior measure Γ_{k-1}^θ, we can use Eq. (2.9) to compute the posterior measure Γ_k^θ. In this way the posterior measure $\{\Gamma_k^\theta\}_{k\in\mathbb{N}^+}$ can be computed recursively. Moreover, using (2.2) again, the conditional likelihood $p(y_k \mid y_{1:k-1}, \theta)$ and the likelihood $p(y_{1:k} \mid \theta)$ can be respectively computed as

$$\begin{aligned}p(y_k \mid y_{1:k-1}, \theta) &= \int p(y_k \mid x_k, y_{1:k-1}, \theta)p(x_k \mid y_{1:k-1}, \theta)\,\mathrm{d}x_k\\ &= \int p(y_k \mid x_k, \theta)p(x_k \mid y_{1:k-1}, \theta)\,\mathrm{d}x_k\end{aligned}$$

$$= (l^\theta_{y_k}, \gamma^\theta_k), \tag{2.10}$$

and

$$\begin{aligned} p(y_{1:k} \mid \theta) &= P(y_1 \mid \theta) \prod_{i=2}^{k} p(y_{t_i} \mid y_{1:i-1}, \theta) \\ &= P(y_1 \mid \theta) \prod_{i=2}^{k} (l^\theta_{y_i}, \gamma^\theta_i)\,. \end{aligned}$$

When (2.1) is a linear Gaussian model, then the Bayesian filter is equivalent to the Kalman filter, which we briefly review in the next subsection.

2.3 Kalman filter

We assume that the state and observations in (2.1) evolve according to a linear Gaussian model. That is the functions f_k and h_k have to take linear forms as follows

$$\begin{aligned} x_k &= F_k x_{k-1} + u_k\,, \\ y_k &= H_k x_k + v_k\,, \qquad k \in \mathbb{N}^+\,, \end{aligned} \tag{2.11}$$

where F_k is a $d \times d$ matrix, H_k is a $m \times d$ matrix and the noise terms u_k (d-dimensional), v_k (m-dimensional) are assumed to be Gaussian with mean 0 and variance Q_k, R_k, respectively. Moreover, the initial state x_0 is assumed to be Gaussian. Due to the Gaussian assumptions and the linear structure of the model in (2.11), one can derive analytic expressions for the prior and posterior measures defined in (2.6) and the algorithm in the *Kalman filter*, see e.g. Chui and Chen (2017), Grewal and Andrews (2015), yields the exact solution to the estimation problem.

Denote by $N(\mathrm{d}x; \mu, \Sigma)$ or $N(\mu, \Sigma)$ the Gaussian distribution with mean μ and Covariance Σ. We also use the generic notation $N(x; \mu, \Sigma)$ to denote the density at x of this normal distribution. Recall from (2.6), the prior and posterior measures and denote by A_{k-1} and P_{k-1} respectively, the mean and the covariance of the posterior measure at time $k-1$. Recall also that we denote θ the vector of all the parameters involved, i.e. those in F_k and H_k for model (2.11). Then the prior measure is given by

$$\gamma^\theta_k(\mathrm{d}x_k) = N(\mathrm{d}x_k; F_k A_{k-1}, F_k P_{k-1} F_k^\top + Q_k)\,,$$

which implies that the prior measure is a conditionally Gaussian measure with mean and covariance respectively given by

$$A_k^- = F_k A_{k-1}\,, \qquad P_k^- = F_k P_{k-1} F_k^\top + Q_k\,.$$

Moreover, the posterior measure is given by

$$\Gamma^\theta_k(\mathrm{d}x_k) = N(\mathrm{d}x_k; A_k, C_k)\,, \tag{2.12}$$

where

$$\begin{aligned} A_k &= A_k^- + P_k^- H_k^\top (H_k P_k^- H_k^\top + R_k)^{-1} (y_k - H_k A_k^-)\,, \\ C_k &= P_k^- - P_k^- H_k^\top (H_k P_k^- H_k^\top + R_k)^{-1} H_k P_k^-\,. \end{aligned}$$

Finally, the conditional likelihood is given by

$$p(y_k \mid y_{1:k-1}, \theta) = N(y_k; H_k A_k^-, H_k P_k^- H_k^\top + R_k)\,. \tag{2.13}$$

Let $S_k = H_k P_k^- H_k^\top + R_k$, $k \in \mathbb{N}^+$. Then we obtain the recursion for the log-likelihood of the observation $\log(p(y_{1:k}) \mid \theta)$ as follows,

$$\begin{aligned}\log(p(y_{1:k}) \mid \theta) = \log p((y_{1:k-1}) \mid \theta)\\ &- \frac{1}{2}\left(m \log 2\pi - \log(\det(S_k))\right)\\ &- \frac{1}{2}(y_k - H_k A_k^-)^\top S_k^{-1}(y_k - H_k A_k^-)\,,\end{aligned}$$

where m is the dimensionality of the data y_k, $k \in \mathbb{N}^+$. Hence, by maximizing the likelihood of the observations, one can determine the optimal parameters of the linear Gaussian system (2.11).

For most non linear non Gaussian models, it is not possible to compute the prior and posterior measures analytically and numerical methods are called for. In this case, the particle filter, which we introduce in the next subsection, is widely used.

2.4 Particle filter

In the *particle filter*, see e.g. Arulampalam et al. (2002), Cappé et al. (2007) and Doucet and Johansen (2009), the prior and posterior distributions are estimated by a Monte Carlo method. With a Monte Carlo method, a certain measure μ is generally estimated by

$$\mu^N(\mathrm{d}x) = \sum_{i=1}^{N} a^{(i)} \delta_{x^{(i)}}(\mathrm{d}x)\,,$$

where $\{x^{(i)}, i = 1, \cdots, N\}$ are i.i.d. random samples from a so-called *importance density* and $\{a^{(i)}, i = 1, \cdots, N\}$ are *the importance weights*. The key part of the particle filter is to choose the importance density and compute the importance weights, see e.g. Doucet (1997). For the general state space model (2.1), suppose the posterior measure Γ_{k-1}^θ at time $k-1$ is estimated by

$$\Gamma_{k-1}^\theta \approx \sum_{i=1}^{N} a_{k-1}^{(i)} \delta_{x_{k-1}^{(i)}}\,.$$

If at time k, the samples $\tilde{x}_k^{(i)}$ are generated from the transition measure $\pi_k^\theta(\mathrm{d}x \mid x_{k-1}^{(i)})$ for $i = 1, \cdots, N$, then using Eq. (2.8), the integral (f, γ_k^θ) can be estimated by

$$(f, \gamma_k^\theta) \approx \sum_{i=1}^{N} a_{k-1}^{(i)} f(\tilde{x}_k^{(i)}).$$

Moreover, using Eq. (2.9), the prior and posterior measures are respectively estimated by

$$\begin{aligned}\gamma_k^\theta &\approx \sum_{i=1}^{N} a_{k-1}^{(i)} \delta_{\tilde{x}_k^{(i)}}\,,\\ \Gamma_k^\theta &\approx \sum_{i=1}^{N} a_k^{(i)} \delta_{\tilde{x}_k^{(i)}}\end{aligned} \tag{2.14}$$

and from (2.10), we deduce the following approximation for the conditional likelihood

$$p(y_k \mid y_{1:k-1}, \theta) \approx \sum_{i=1}^{N} a_{k-1}^{(i)} l_{y_k}^{\theta}(\tilde{x}_k^{(i)}) . \tag{2.15}$$

Consequently, the integral (f, Γ_k^{θ}) can be estimated by

$$\begin{aligned}(f, \Gamma_k^{\theta}) &\approx \frac{\sum_{i=1}^{N} a_{k-1}^{(i)} l_{y_k}^{\theta}(\tilde{x}_k^{(i)}) f(\tilde{x}_k^{(i)})}{\sum_{i=1}^{N} a_{k-1}^{(i)} l_{y_k}^{\theta}(\tilde{x}_k^{(i)})} \\ &= \sum_{i=1}^{N} a_k^{(i)} f(\tilde{x}_k^{(i)}) ,\end{aligned}$$

where the weights $a_k^{(i)}$ are defined by

$$a_k^{(i)} = \frac{a_{k-1}^{(i)} l_{y_k}^{\theta}(\tilde{x}_k^{(i)})}{\sum_{i=1}^{N} a_{k-1}^{(i)} l_{y_k}^{\theta}(\tilde{x}_k^{(i)})} . \tag{2.16}$$

Eqs. (2.14) and (2.15) show how to sequentially estimate the posterior measure Γ_k^{θ} using the Monte Carlo method. This type of particle filter is often referred to as *sequential particle filter*. It is a specific member of the family termed the bootstrap particle filter (see Gordon et al. 1993). In Doucet (1997) it is shown that the variance of the importance weights decreases stochastically over time. This will lead the importance weights to be concentrated on a small amount of sampled particles. This problem is called *degeneracy*. To address the rapid degeneracy problem, the *sampling-importance resampling (SIR)* method, see e.g. Doucet (1997), Pitt and Shephard (1999), is introduced to eliminate the samples with low importance weight and multiply the samples with high importance weight. In SIR, once the approximation of the posterior measure $\Gamma_k^{\theta} \approx \sum_{i=1}^{N} a_k^{(i)} \delta_{\tilde{x}_k^{(i)}}$ is obtained, new, re-sampled, particles $x_k^{(j)}$ are i.i.d. sampled from this approximated measure, i.e. every $x_k^{(j)}$ is independently chosen from the $\tilde{x}_k^{(i)}$ with probabilities $a_k^{(i)}$, for $i = 1, \cdots, N$. This step can be accomplished by sampling integers j from $\{1, \ldots, n\}$ with probabilities $a_k^{(i)}$, $i = 1, \cdots, N$. Then the new estimation on the posterior measure Γ_k^{θ} is given by

$$\Gamma_k^{\theta} \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{x_k^{(i)}}$$

and the new estimate of the conditional likelihood is

$$p(y_k \mid y_{1:k-1}, \theta) \approx \frac{1}{N} \sum_{i=1}^{N} l_{y_k}^{\theta}(x_k^{(i)}) .$$

2.5 Static model parameters estimation using particle filter

When the parameters are known, the particle filter is a quite effective algorithm for latent variable estimation. However, if the parameters are not known beforehand, it is a very challenging task to estimate the parameters and the latent states using the particle filter. Here we take a Bayesian approach to estimate the parameters. The estimation of the parameters in online estimation requires the computation of the posterior distribution of θ, i.e., $p(\theta \mid y_{1:k})$, $k \in \mathbb{N}^+$. Using Bayes' rule, one can represent the posterior density as

$$p(\theta \mid y_{1:k}) = \frac{p(y_k \mid y_{1:k-1}, \theta)p(\theta \mid y_{1:k-1})}{\int p(y_k \mid y_{1:k-1}, \theta)p(\theta \mid y_{1:k-1})\,\mathrm{d}\theta}\,.$$

To estimate the density of θ given $y_{1:k}$, a straightforward way is to sample parameter particles from the former posterior distribution $p(\theta \mid y_{1:k-1})$. Denote the samples by $\{\theta^{(i)}, i = 1, \cdots, N\}$, then the measure $p(\mathrm{d}\theta \mid y_{1:k})$ at time k can be approximated by

$$\sum_{i=1}^{N} \frac{p(y_k \mid y_{1:k-1}, \theta^{(i)})}{\sum_{i=1}^{N} p(y_k \mid y_{1:k-1}, \theta^{(i)})}\delta_{\theta^{(i)}}(\mathrm{d}\theta) = \sum_{i=1}^{N} w_k^{\theta^{(i)}} \delta_{\theta^{(i)}}(\mathrm{d}\theta)\,, \tag{2.17}$$

where the weights $w_k^{\theta^{(i)}}$, $i = 1, \cdots N$, are defined by

$$w_k^{\theta^{(i)}} = \frac{p(y_k \mid y_{1:k-1}, \theta^{(i)})}{\sum_{i=1}^{N} p(y_k \mid y_{1:k-1}, \theta^{(i)})}\,. \tag{2.18}$$

There are two issues to implement (2.17). One is that sampling from the former posterior distribution $p(\theta \mid y_{1:k-1})$ usually cannot be carried out exactly. Another is that often the likelihood $p(y_k \mid y_{1:k-1}, \theta^{(i)})$ cannot be computed theoretically. These two latter issues can be tackled by using the *recursive nested particle filter (RNPF)*, recently introduced in Crisan and Míguez (2018), which is presented below.

2.5.1 Recursive nested particle filter

In the RNPF, a two layer Monte Carlo method is used. In the first layer, also referred to as outer layer, new parameter samples are generated by using a kernel function. This step is usually called *jittering* and the kernel is referred to as *the jittering kernel*. In the second layer, also called inner layer, a particle filter is applied to approximate the conditional likelihood $p(y_k \mid y_{1:k-1}, \theta^{(i)})$. In the following paragraph of this section we present the RNPF in more detail and introduce its ensuing Algorithm 2.1.

First, assume that θ has a compact support $D_\theta \subset \mathbb{R}^{d_\theta}$, where d_θ is the dimension of θ. Moreover assume at time $k-1$, one can generate a random grid of samples in the parameter space D_θ, say $\{\theta_{k-1}^{(i)}, i = 1, \cdots, N\}$, and for each $\theta_{k-1}^{(i)}$, we have the set of particles in the state space $\{x_{k-1}^{(i,j)}, 1 \leq j \leq M\}$.

- **Jittering.** Given the parameters samples $\{\theta_{k-1}^{(i)}, i = 1, \cdots, N\}$ at time $k-1$, new particles $\{\tilde{\theta}_k^{(i)}, i = 1, \cdots, N\}$ at time k are generated by some Markov kernels denoted by $\kappa(d\theta \mid \theta_{k-1}^{(i)}) : \mathcal{B}(D_\theta) \times D_\theta \to [0, 1]$ (step 1.a in Algorithm 2.1 below). This step is the outer Monte Carlo layer.
- **Update.** From Eqs. (2.8) and (2.10), we know that for a given $\tilde{\theta}$, the likelihood function is obtained by calculating the integral

$$p(y_k \mid y_{1:k-1}, \tilde{\theta}) = ((l_{y_k}^{\tilde{\theta}}, \pi_k^{\tilde{\theta}}), \Gamma_{k-1}^{\tilde{\theta}})\,.$$

In order to compute this latter integral, the posterior measure at time $k-1$, $\Gamma_{k-1}^{\tilde{\theta}}$, needs to be known. In the standard Bayesian filter, the parameters are fixed over time and this posterior measure is computed at time $k-1$ by using Eq. (2.9). However in this case, this measure is not directly available since the parameter has evolved from θ at time $k-1$ to $\tilde{\theta}$ at time k. In order to compute $\Gamma_{k-1}^{\tilde{\theta}}$, one needs to re-run a filter from time 1 to k, which makes the algorithm not recursive and very time consuming. The authors in Crisan and Míguez (2018) solved this latter problem by assuming that Γ_{k-1}^{θ} is continuous w.r.t. $\theta \in D_\theta$, which means that when $\theta \approx \tilde{\theta}$, then $\Gamma_{k-1}^{\theta} \approx \Gamma_{k-1}^{\tilde{\theta}}$. Therefore by considering a rather small variance in the jittering kernel, one can use the particle approximation of the filter computed for θ at time $k-1$ as a particle approximation of the filter for the new sampled $\tilde{\theta}$ at time k.

In the RNPF, the jittering kernel is chosen such that the mutation step from θ_{k-1} to $\tilde{\theta}_k$ is sufficiently small, see Sect. 4.2 in Crisan and Míguez (2018). Then for each $\tilde{\theta}_k^{(i)}$, $\{i = 1, \cdots, N\}$, a sequential nested particle filter (see Sect. 2.4 for the description of the particle filter methodology) is used for the state space to obtain $\{\tilde{x}_k^{(i,j)},\ 1 \leq j \leq M\}$; see steps 1.b, 1.d, 1.e in Algorithm 2.1 below. This is the inner Monte Carlo layer.

– **Resampling.** The outer layer Monte Carlo method in the update step above provides an approximation of the likelihood $p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})$, $i = 1, \cdots, N$, (step 1.c in Algorithm 2.1) which are used to re-weight the parameter particles and obtain $\{\theta_k^{(i)}, x_k^{(i,j)},\ i = 1 \cdots, N,\ j = 1, \cdots, M\}$; see step 2 in Algorithm 2.1.

The RNPF is introduced in Crisan and Míguez (2018). We reproduce it here for the sake of completeness.

Algorithm 2.1 **(sequential nested particle filter for parameter estimation)**

Initialization: Assume an initial distribution $p(\theta_0)$ for the parameters and $p(x_0)$ for the states, and sample from the initial distributions to get N particles $\{\theta_0^{(i)}, i = 1, \cdots, N\}$ and $N \times M$ particles $\{x_0^{(i,j)}, i = 1, \cdots, N,\ j = 1, \cdots, M\}$.

Recursion:

1. Filtering: given $\{\theta_{k-1}^{(i)}, x_{k-1}^{(i,j)}\}$, for each $i = 1, \cdots, N$,
    a. (jittering, outer Monte Carlo layer) sample new parameters $\tilde{\theta}_k^{(i)}$ from the jittering kernel $\kappa(d\theta \mid \theta_{k-1}^{(i)})$,
    b. (together with the next two steps, this is the update part) sample new states $\hat{x}_k^{(i,j)}$, $j = 1, \cdots, M$, from the transition measure $\pi_k^{\tilde{\theta}_k^{(i)}}(\mathrm{d}x \mid x_{k-1}^{(i,j)})$ (inner Monte Carlo Layer),
    c. compute $p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)}) \approx \frac{1}{M}\sum_{j=1}^{M} l_{y_k}^{\tilde{\theta}_k^{(i)}}(\hat{x}_k^{(i,j)})$,
    d. compute the weights for the state space using Eq. (2.16)

    $$a_k^{(j)} = \frac{a_{k-1}^{(j)} p(y_k \mid \hat{x}_k^{(i,j)}, \tilde{\theta}_k^{(i)})}{\sum_{j=1}^{M} a_{k-1}^{(j)} p(y_k \mid \hat{x}_k^{(i,j)}, \tilde{\theta}_k^{(i)})}, \qquad j = 1, \cdots, M\,,$$

    e. resample the $\hat{x}^{(i,p)}$: set $\tilde{x}^{(i,j)}$ equal to $\hat{x}^{(i,p)}$ with probability $a_k^{(p)}$, where $j, p \in \{1, \cdots M\}$.
2. Resampling of the $\{\theta_k^{(i)}\}$: compute the weights for the parameters space using Eq. (2.18)

$$w_k^{\tilde{\theta}_k^{(i)}} = \frac{p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})}{\sum_{i=1}^{N} p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})}, \qquad i = 1, \cdots, N\,. \tag{2.19}$$

For $i = 1, \cdots, N$, set $\{\theta_k^{(i)}, x_k^{(i,j)}\}_{1\leq j\leq M}$ equal to $\{\tilde{\theta}_k^{(p)}, \tilde{x}_k^{(p,j)}\}_{1\leq j\leq M}$ with probability $w_k^{\tilde{\theta}_k^{(p)}}$, where $p \in \{1, \cdots, N\}$.
3. Go back to the filtering step.

2.5.2 A note on the convergence of the RNPF

A convergence study of Algorithm 2.1 was carried out in Lemmas 3 to 6 and Theorems 2 and 3 in Crisan and Míguez (2018), where the reasoning was split in the three steps of the algorithm: the jittering, the update and the resampling. In this latter paper, it was proven that, under some regularity conditions, the L^p-norms of the approximation errors, induced by these different steps, vanish with rate proportional to $\frac{1}{\sqrt{N}}$ and $\frac{1}{\sqrt{M}}$. Recall here that N and $N \times M$ are respectively the number of samples in the parameter space and the number of particles in the state space. A similar result was proven for the approximation of the joint posterior distribution of the parameters and the state variables. We will make use of some of these convergence results later in Sect. 5 to carry out convergence study of our proposed algorithm, Algorithm 4.2.

Under the assumption that the posterior measure $\Gamma_k^\theta(\mathrm{d}x)$ is continuous w.r.t. the parameter θ and when the mutation step of the parameters is small enough, the RNPF is a recursive algorithm. This makes the RNPF more efficient than non-recursive methods such as *sequential Monte Carlo square*, see Chopin et al. (2013), and *Markov Chain Monte Carlo* methods, see Gamerman and Lopes (2006), Geweke and Tanizaki (1999) and Higdon (1998). The drawbacks of the RNPF are its heavy computational burden and slow convergence speed which are respectively due to the nested simulations in the two Monte Carlo layers and the small mutation step of the parameters. In many applications, including some in financial modeling, the time length of the data is quite limited, and often too short to observe convergent behavior of the RNPF. To tackle this problem, we propose a new methodology in Sect. 4.

3 Parameters estimation in short rate models

Here we present a rather general model, an affine process, particularly relevant in mathematical finance for instance where one is interested in estimating the parameters of the short rate curve given the observed data. It motivates the kind of system that we will consider and to which the new (Kalman particle) filter of Sect. 4 will be applied.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space satisfying the usual conditions and $(W_t)_{t\geq 0}$ be a d-dimensional Brownian motion. In this paper, although our results can be applied to general state space models of type (2.1), we will mainly consider dynamics of the type

$$\mathrm{d}x_t = A(\beta - x_t)\,\mathrm{d}t + \left(\Sigma + \tilde{\Sigma}\sqrt{x_t^{(1)}}\right)\mathrm{d}W_t,\; x_0 = x \in \mathbb{R}^d\,, \tag{3.1}$$

where A, Σ and $\tilde{\Sigma}$ are $d \times d$-matrices, β is a d-vector and its first component is non-negative, and $(x_t^{(1)})_{t\geq 0}$ is the first component of $(x_t)_{t\geq 0}$. We assume the matrix A is diagonal and we denote the diagonal elements of A by $\alpha_1, \cdots, \alpha_d$. Consider some integers $p, q \geq 0$ with $p + q = d$. When $\Sigma\tilde{\Sigma} = \tilde{\Sigma}\Sigma = 0$ (below we specialise to the cases $\Sigma = 0$ or $\tilde{\Sigma} = 0$) and the parameters of the model (3.1) satisfy certain conditions known in the literature as *admissibility conditions*, the process $(x_t)_{t\geq 0}$ is $\mathbb{R}_+^p \times \mathbb{R}^q$-valued *affine* process, see Duffie et al. (2003), Duffie et al. (2000), Keller-Ressel and Mayerhofer (2015)) for an overview of affine processes. In “Appendix A.1” we specify the admissibility of the parameters of the dynamics (3.1). In our context, the short rate evolution will be described by a process $(r_t)_{t\geq 0}$ given in terms of $(x_t)_{t\geq 0}$ by

$$r_t = c + \gamma^\top x_t \,,$$

where $c \in \mathbb{R}$, $\gamma \in \mathbb{R}^d$. Let $T > 0$ be the maturity time, then the *zero coupon bond* price at time $t < T$ is defined as

$$P(t, T) = \mathbb{E}[\mathrm{e}^{-\int_t^T r(s)\,\mathrm{d}s} \mid \mathcal{F}_t]\,,$$

and the corresponding zero rates, also called yields, are defined as

$$-\log P(t, T)/(T - t)\,.$$

The fact that the process $(x_t)_{t\geq 0}$ is affine, which happens if $\Sigma\tilde{\Sigma}$ is zero, allows one to obtain an explicit formula for the zero coupon bond price, i.e.

$$P(t, T) = \mathrm{e}^{-\phi(T-t,0)-\psi(T-t,0)^\top x(t)}\,. \tag{3.2}$$

The functions ϕ and ψ are the solutions to some ordinary differential equations, which are often referred to as the Riccati equations, see Theorem A.1 in “Appendix A.2” for details. Then, if $\tilde{\Sigma} = 0$ (first case), Eq. (3.2) holds for (ϕ, ψ) the solution to (A.1). If $\Sigma = 0$ (second case), then (3.2) holds for (ϕ, ψ) the solution to (A.2). Denote the time to maturity $T - t$ by τ, then the zero rate at time t with time to maturity τ can be computed by $\frac{1}{\tau} R_t(\tau)$, where

$$R_t(\tau) := \phi(\tau, 0) + \psi(\tau, 0)^\top x_t. \tag{3.3}$$

In the market, we can obtain the data for zero rates at discrete-time instants t_k with certain times to maturity $\tau_1, \cdots, \tau_L$, call these data $R_k(\tau_l)$. We believe these data contain noise, hence at time k we observe

$$y_k = [y_k(\tau_1), \cdots, y_k(\tau_L)] = [R_k(\tau_1), \cdots, R_k(\tau_L)]^\top + v_k,$$

for $k = 1, \cdots, K$, and v_k is an L-dimensional random vector which presents the noise in the observed data. Let $0 = t_0 \leq t_1, \cdots, t_n = T$ be a partition of the time interval $[0, T]$. Then, considering a time-discrete version $x_k := x_{t_k}$, $k \in \mathbb{N}^+$, of the affine process $(x_t)_{t\geq 0}$, our aim is to derive the parameters of the latent state process $(x_k)_{k\in\mathbb{N}^+}$ given the observations $(y_k)_{k\in\mathbb{N}^+}$. To be more precise, we consider the following state space model, the observation equation can be seen as of the general form in (2.11) by enlarging the state vector,

$$\begin{aligned} x_k = {} & \mathrm{e}^{-A(t_k-t_{k-1})} x_{k-1} + \left(I - \mathrm{e}^{-A(t_k-t_{k-1})}\right)\beta \\ & + \int_{t_{k-1}}^{t_k} \mathrm{e}^{-A(t_k-u)} \left(\Sigma + \tilde{\Sigma}\sqrt{x_u^{(1)}}\right) \mathrm{d}W_u\,, \end{aligned} \tag{3.4}$$

$$y_k = H_k x_k + H_k^0 + v_k\,, \tag{3.5}$$

where I is the identity matrix, $(x_k)_{k\in\mathbb{N}^+}$ is the latent process, $(y_k)_{k\in\mathbb{N}^+}$ represents the observations, H_k is a $L \times d$ matrix with each row equal to $-\psi(\tau_l, 0)/\tau_l$, $l = 1, \cdots, L$, H_k^0 is the column vector $[-\phi(\tau_1, 0)/\tau_1, \cdots, -\phi(\tau_L, 0)/\tau_L]^\top$ and v_k represents the noise. Note that the x_k from (3.4) forms a discrete-time sample from the continuous process that solves (3.1) and that (3.4) exactly represents x_t at time $t = t_k$, it is not a recipe for an approximation.

The aim is to estimate the model parameters A, β, Σ, $\tilde{\Sigma}$ given the observation vector $y_{1:k}$ and the variance of v_k. As usual, in the sequel we collectively denote these parameters by θ.

We end this section by giving some examples of the models of type (3.1) which are well known in the literature and to which we return with numerical experiments in Sect. 6. For $d = 1$, $p = 1$, $\Sigma = 0$, $\tilde{\Sigma} \neq 0$, one obtains the *Cox-Ingersoll-Ross* (CIR) model, see Cox et al. (1985), i.e.,

$$\mathrm{d}x_t = \alpha_1(\beta - x_t)\,\mathrm{d}t + \tilde{\Sigma}\sqrt{x_t}\,\mathrm{d}W_t\,. \tag{3.6}$$

For $d = 2$, $p = 0$, $\beta^{(1)} = \beta^{(2)} = 0$, $\Sigma \neq 0$, $\tilde{\Sigma} = 0$, one obtains the two-factor *Hull-White* model with constant parameters (also known as the two-factor *Vasiček* model), with mean-reversion level 0, see Hull and White (1990), i.e.,

$$\begin{aligned} \mathrm{d}x_t^{(1)} &= -\alpha_{11}x_t^{(1)}\,\mathrm{d}t + \Sigma_{11}\,\mathrm{d}W_t^{(1)} + \Sigma_{12}\,\mathrm{d}W_t^{(2)}\,,\\ \mathrm{d}x_t^{(2)} &= -\alpha_{22}x_t^{(2)}\,\mathrm{d}t + \Sigma_{21}\,\mathrm{d}W_t^{(1)} + \Sigma_{22}\,\mathrm{d}W_t^{(2)}\,. \end{aligned} \tag{3.7}$$

For $d = 2$, $p = q = 1$, $\Sigma = 0$, one obtains the *stochastic volatility* model, see Heston (1993), in which the first component, $(x_t^{(1)})_{t\geq 0}$, represents the stochastic volatility of the short rate $(x_t^{(2)})_{t\geq 0}$, i.e.,

$$\begin{aligned} \mathrm{d}x_t^{(1)} &= \alpha_{11}(\beta_1 - x_t^{(1)})\,\mathrm{d}t + \sqrt{x_t^{(1)}}\left(\tilde{\Sigma}_{11}\,\mathrm{d}W_t^{(1)} + \tilde{\Sigma}_{12}\,\mathrm{d}W_t^{(2)}\right)\,,\\ \mathrm{d}x_t^{(2)} &= \alpha_{22}(\beta_2 - x_t^{(2)})\,\mathrm{d}t + \sqrt{x_t^{(1)}}\left(\tilde{\Sigma}_{21}\,\mathrm{d}W_t^{(1)} + \tilde{\Sigma}_{22}\,\mathrm{d}W_t^{(2)}\right)\,. \end{aligned} \tag{3.8}$$

4 Kalman particle filter for online parameters estimation

In this section, we introduce the *Kalman particle filter* for online parameter estimation. It is a semi-recursive algorithm that combines the Kalman filter and the particle filter. In this new approach, we consider a two layers method as in the RNPF algorithm. In the outer layer, we sample the particles of the model parameters using some Markovian Gaussian kernel which is updated at each time step. In the inner layer, the distribution of the state process and the conditional likelihood function $p(y_k \mid y_{1:k-1}, \theta_k^{(i)})$, which is used to re-weight the parameter particles in the outer layer, are estimated given the sampled parameter particles.

There are two main differences between our proposed Kalman particle filter algorithm and the RNPF algorithm. The first difference is that in the outer layer we use dynamic jittering functions, i.e. the jittering functions can change over time. Specially, in this paper we specify two jittering functions to sample the model parameters, see (4.5) and (4.9) as described in Sect. 4.1 below. The second difference is that we use the Kalman filter, instead of the particle filter, to update the underlying states in the inner layer. Note that in case the state space does not follow linear Gaussian dynamics, the literature offers different alternatives, see Bruno (2013) for a Monte Carlo approach, or the *Gaussian mixture*, see Sorenson and Alspach (1971), or Kalman filter extensions such as the *extended Kalman filter*, see Einicke and White (1999) and Wan and Nelson (2002), the *unscented Kalman filter*, see Wan and van der Merwe (2002) . In these latter methodologies, the idea is to consider an approximation of the state variables which is linear and Gaussian, and then run a Kalman filter on the approximation. When the model is not Gaussian, such an approximation introduces bias. In Sect. 5, we will carry a convergence analysis of our algorithm and we will prove that the bias induced by the Gaussian approximation of the model (3.4), (3.5) indeed vanishes when the time step tends to zero.

The use of the two jittering functions in the outer layer and of the Kalman filter in the inner layer allows us to obtain an algorithm that has faster convergence speed and less computational complexity than the RNPF algorithm. This will be further illustrated in the examples in Sect. 6.

4.1 Static model estimation

As described in Sect. 2.5, in order to sequentially estimate the posterior density $p(\theta \mid y_{1:k})$, $k = 1, \cdots, K$, we face two issues: how to sample particles from the former posterior distribution $p(\theta \mid y_{1:k-1})$ and how to compute the conditional likelihood $p(y_k \mid y_{1:k-1}, \theta)$. First, we consider the sampling problem and we introduce the first jittering kernel that we use to update the parameter space θ. We assume that the parameter θ has a compact domain $D_\theta \subset \mathbb{R}^{d_\theta}$, with d_θ the dimension of θ. For the time being, θ is an abstract parameter, which will be specified later when a specific model is assumed.

4.1.1 Truncated Gaussian kernel with changing covariance

Recall that we use the generic notation $N(x; \mu, \Sigma)$ to denote the density at x of the normal distribution with mean vector μ and covariance matrix Σ. We choose a Gaussian kernel such that the conditional density of θ_k is given as

$$p(\theta_k \mid \theta_{k-1}) = N(\theta_k; \mu(\theta_{k-1}), \Sigma(\theta_{k-1}))\,,$$

with $\mu(\theta_{k-1})$ and $\Sigma(\theta_{k-1})$ being respectively the conditional mean and covariance of θ_k given θ_{k-1}. Then if the parameters θ_k are jittered from this Gaussian kernel, one can easily derive that

$$\begin{aligned} \mathbb{E}(\theta_k) &= \mathbb{E}(\mu(\theta_{k-1})),\\ \mathrm{Var}(\theta_k) &= \mathbb{E}(\Sigma(\theta_{k-1})) + \mathrm{Var}(\mu(\theta_{k-1}))\,. \end{aligned} \tag{4.1}$$

Ideally, the jittering should not introduce bias and information loss (artificial increase in the variance), see Liu and West (2001), which means that $\mathbb{E}(\theta_k) = \mathbb{E}(\theta_{k-1})$ and $\mathrm{Var}(\theta_k) = \mathrm{Var}(\theta_{k-1})$, $k = 1, \cdots, K$. The latter, together with Eqs. (4.1) imply

$$\begin{aligned} \mathbb{E}(\mu(\theta_{k-1})) &= \mathbb{E}(\theta_{k-1}),\\ \mathbb{E}(\Sigma(\theta_{k-1})) + \mathrm{Var}(\mu(\theta_{k-1})) &= \mathrm{Var}(\theta_{k-1})\,. \end{aligned} \tag{4.2}$$

To achieve that, the Liu and West (2001) applies a shrinkage to the kernel. We will apply the same technique although the jittering function is used differently in our case. If one assumes a deterministic jittering covariance, i.e. $\Sigma(\theta_{k-1}) = \Sigma_{k-1}$ and a linear mean function

$$\mu(\theta_{k-1}) = a\theta_{k-1} + (1-a)\mathbb{E}(\theta_{k-1}), \text{ for some } a \in (0, 1), \tag{4.3}$$

then the jittering kernel satisfying (4.2) is given by

$$p(\theta_k \mid \theta_{k-1}) = N(\theta_k; \mu(\theta_{k-1}), \Sigma(\theta_{k-1}))\,, \tag{4.4}$$

where $\Sigma(\theta_{k-1}) = (1-a^2)\mathrm{Var}(\theta_{k-1})$. The kernel in (4.4) is the same jittering kernel as used in the Liu and West (2001). We will refer to the number a as the discount factor. However, note that the domain of the kernel (4.4) is not compact, whereas we assumed the domain of θ to be compact. So instead of directly using the kernel (4.4), we truncate this Gaussian distribution to the compact domain D_θ. The truncated kernel is denoted

$$p(\theta_k \mid \theta_{k-1}) = \text{TruncNorm}(\theta_k; \mu(\theta_{k-1}), \Sigma(\theta_{k-1}), D_\theta)\,, \tag{4.5}$$

with $\text{TruncNorm}(x; \mu, \Sigma, D_\theta)$ the truncated normal distribution which truncates the normal distribution with mean μ and covariance Σ to the range D_θ.

Before we present our methodology for jittering in detail, we introduce the following assumption which we need in our recursive algorithm later. We will use this assumption in Sect. 5 to prove the convergence of our proposed algorithm.

Assumption 4.1 The jittering kernels $\kappa_k^N(d\theta \mid \theta') : \mathcal{B}(D_\theta) \times D_\theta \to [0, 1]$, for $k = 1, \cdots, N$, and θ' taking values in a compact set $D_\theta \subset \mathbb{R}^d$, satisfy the following inequalities

$$\sup_{\theta' \in D_\theta} \int \left| f(\theta) - f(\theta') \right| \kappa_k^N(d\theta \mid \theta') \leq \frac{e_{1,k}\, \|f\|_\infty}{\sqrt{N}}\,, \tag{4.6}$$

for all bounded function $f : D_\theta \to \mathbb{R}$ and some positive constant $e_{1,k}$ which is independent of f, and

$$\sup_{\theta' \in D_\theta} \int \left\| \theta - \theta' \right\|^p \kappa_k^N(d\theta \mid \theta') \leq \frac{e_{2,k}^p}{\sqrt{N^p}}\,, \tag{4.7}$$

for $p \geq 1$ and some positive constant $e_{2,k}$.

Let $f : D_\theta \to \mathbb{R}$ be a bounded Lipschitz function. In Proposition 1 of Appendix C in Crisan and Míguez (2018), set $\epsilon_n = 1$ there, it is shown that if for any $p \geq 1$, the jittering kernels $\kappa_k^N(\mathrm{d}\theta \mid \theta')$ satisfy

$$\sup_{\theta' \in D_\theta} \int \left\| \theta - \theta' \right\|^2 \kappa_k^N(\mathrm{d}\theta \mid \theta') \leq \frac{c}{\sqrt{N^{p+2}}}\,, \tag{4.8}$$

for some positive constant c independent of N, then Assumption 4.1 holds.

The jittering kernels of type (4.5) have an appealing property, the covariance can change over time. This aspect helps us to design an algorithm with the following attractive feature. Initially, since we lack information on the unknown parameters, a larger covariance can lead to a faster convergence of the parameters to the high likelihood area. Over time, the filter refines the estimate of the fixed parameters until at some points a very small variance has been reached which makes the parameter estimation more accurate. However, a direct application of this kernel does not yield a recursive method since it generally does not satisfy Assumption 4.1. Hence, it is unclear whether the algorithm converges. To tackle this issue, we introduce a second Gaussian jittering kernel which satisfies Assumption 4.1. This second jittering kernel will allow us to obtain a recursive method. The details will be described below in Sect. 4.1.2.

Remark 4.1 Although in this paper we have specified the use of (truncated) Gaussian kernels with a certain mean and variance, the dynamic kernel set up is very generic. In implementation, one can freely choose another kernel that fits its purpose. For example, one can define the variance of the jittering kernel as a monotonically decreasing function of time so that the convergence speed can be manually controlled.

4.1.2 Description of the Kalman particle filter methodology

Here we present our methodology to Gaussian and linear models. When the model is not Gaussian and linear, we approximate it by a Gaussian linear model and hence we can follow the same methodology as described below to the approximation. Before we present the algorithms, we introduce some notation and convention. We denote by $\mathrm{diag}(A)$ the diagonal matrix that has the same diagonal elements as in a given matrix A. With a diagonal matrix V, $\mathrm{diag}(A) < V$ means every diagonal element of $\mathrm{diag}(A)$ is smaller than the corresponding diagonal element in V. Similarly, an expression like $\min\{\mathrm{diag}(A), V\}$ is to be interpreted as element wise. These notational conventions apply to all algorithms below.

In the *recursive step* explained below we set thresholds on the covariance matrices of the jittering kernel so that Assumption 4.1 holds. Notice that although the (co)variance matrices used in the jittering kernel of the recursive step do not have to be diagonal when Assumption 4.1 holds, they are chosen to be diagonal for the convenience of the implementation. Hence, in the jittering kernel in (4.9) below of the recursive step we use $\mathrm{diag}(\varSigma(\theta_k))$ instead of $\varSigma(\theta_k)$.

The non recursive step. Assume at time $k-1$, one can generate a random grid of samples in the parameter space, say $\{\theta_{k-1}^{(i)}, i = 1, \cdots, N\}$.

- **Jittering step 1.** Here we apply the kernel (4.5), referred to as jittering kernel 1, to obtain new samples $\{\tilde{\theta}_k^{(i)}, i = 1, \cdots, N\}$ (step 1.a.i in Algorithm 4.2 below).
- **Update.** In order to compute the posterior measure $\varGamma_k^{\tilde{\theta}_k^{(i)}}$ at time k, one needs to know the mean $B_{k-1}^{\tilde{\theta}_k^{(i)}}$ and the covariance $P_{k-1}^{\tilde{\theta}_k^{(i)}}$ at time $k-1$ of the posterior distribution, see Formula (2.12) and note that we make the dependence on $\tilde{\theta}_k^{(i)}$ clear in the notation. However, these latter quantities are not available since the parameter has evolved from θ_{k-1} at time k to $\tilde{\theta}_k$ at time k, see also the discussion in Sect. 2.5. Hence at this step, the algorithm does not run recursively and at every time k where a new parameter particle is sampled, the inner filter re-runs from time t_0 to t_k (step 1.a.ii in Algorithm 4.2). Moreover, the conditional likelihood function $p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)}), i = 1, \cdots, N$ is computed in the inner filter, using Eq. (2.13). This latter will be used to re-weight the parameter particles, see step 1.c in Algorithm 4.2 below.
- **Resampling.** We use a resampling technique to obtain $\{\theta_k^{(i)}, B_k^{(i)}, P_k^{(i)}, i = 1, \cdots, N\}$, further specified in step 2 of Algorithm 4.2.

The recursive step. In this step we first need to specify a diagonal matrix V_N. This chosen V_N serves as the threshold for the (co)variance of the jittering kernel satisfying Assumption 4.1. For example, the i-th diagonal element of V_N can be defined as $\frac{c_i}{\sqrt{N^{p+2}}}$ with arbitrary positive constants c_i. Once at some time point t_l, $\mathrm{diag}(\varSigma(\theta_l)) = (1-a^2)$ $\mathrm{diag}(\mathrm{Var}(\theta_l)) < V_N$, we switch to the jittering kernel 2. In a practical implementation, one could also set a floor, a diagonal matrix V_f, with $0 \leq V_f \leq V_N$, for the elements of the jittering variance matrix to prevent the algorithm of getting stuck.

- **Jittering step 2.** We apply the jittering kernel 2,

$$p(\theta_k^{(i)} \mid \theta_{k-1}^{(i)}) = \mathrm{TruncNorm}(\theta_k^{(i)}; \theta_{k-1}^{(i)}, \min\{\max\{\mathrm{diag}(\varSigma(\theta_l)), V_f\}, V_N\}, D_\theta)\,, \tag{4.9}$$

for $k \geq l+1$, see step 1.b.i of Algorithm 4.2.

- **Update.** From time t_{l+1} on, we have a recursive algorithm based on the idea to approximate the posterior measure $\Gamma_k^{\tilde{\theta}_k^{(i)}}$ by $\hat{\Gamma}_k^{\theta_{k-1}^{(i)}}$, which is computed using $\tilde{\theta}_k^{(i)}$, $B_{k-1}^{\theta_{k-1}^{(i)}}$ and the covariance $P_{k-1}^{\theta_{k-1}^{(i)}}$ (step 1.a.ii in Algorithm 4.2). Note that here we use the Kalman filter. Note that, as in Sect. 2.5.1, here we assume that Γ_k^{θ} is continuous w.r.t. $\theta \in D_\theta$. Moreover, the marginal likelihood $p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})$, $i = 1, \cdots, N$ is approximated by the inner filter using Eq. (2.13), step 1.c in Algorithm 4.2. This latter will be used to re-weight the parameter particles
- **Resampling.** Apply a resampling technique to obtain $\{\theta_k^{(i)}, B_k^{(i)}, P_k^{(i)}, i = 1, \cdots, N\}$, see step 2 of Algorithm 4.2.

We now introduce the Kalman particle algorithm. Recall $\mu(\theta_{k-1})$ and $\Sigma(\theta_{k-1})$ from (4.5) and the discount factor a from (4.3).

Algorithm 4.2 (**Kalman particle filter for static parameter model**)
Initialization:

1. set the number of particles N, a value for the discounting factor a, a switching variance level V_N and a floored variance level V_f,
2. assume an initial distribution $p(\theta_0)$ for the parameters,
3. sample from the initial distribution to get N particles $\{\theta_0^{(i)}, i = 1, \cdots, N\}$ for the parameters,
4. for each particle $\theta_0^{(i)}$, assign the same initial mean $B_0^{(i)}$ and covariance value $P_0^{(i)}$ of the posterior distribution $\Gamma_0^{\theta_0^{(i)}}$, $i = 1, \cdots, N$.

Recursion:

1. Filtering: given $\{\theta_{k-1}^{(i)}, i = 1, \cdots, N\}$,
   a. as long as $\operatorname{diag}(\Sigma(\theta_{k-1}))$ is not smaller than V_N (jittering case 1), for each $i = 1, \cdots, N$,
      i. sample new parameters from the kernel (4.5), i.e.
$$\tilde{\theta}_k^{(i)} \sim \text{TruncNorm}\left(\tilde{\theta}_k; \mu(\theta_{k-1}^{(i)}), \Sigma(\theta_{k-1}), D_\theta\right),$$
      ii. based on the parameter $\tilde{\theta}_k^{(i)}$, use the Kalman filter to compute the mean and covariance of the posterior distribution from time 1 to k and hence obtain $\Gamma_k^{\tilde{\theta}_k^{(i)}}$,
   b. once $\operatorname{diag}(\Sigma(\theta_{l-1})) < V_N$ (jittering case 2), for some t_l, then for $k \geq l + 1$ and $i = 1, \cdots, N$, given $\{B_{k-1}^{(i)}, P_{k-1}^{(i)}\}$,
      i. sample new parameters from the kernel (4.9), i.e.
$$\tilde{\theta}_k^{(i)} \sim \text{TruncNorm}(\tilde{\theta}_k; \theta_{k-1}^{(i)}, \min\{\max\{\operatorname{diag}(\Sigma(\theta_{k-1})), V_f\}, V_N\}, D_\theta),$$
      ii. based on the parameters $\tilde{\theta}_k^{(i)}$, $B_{k-1}^{(i)}$ and $P_{k-1}^{(i)}$, use the Kalman filter to compute the mean $\tilde{B}_{k-1}^{(i)}$ and covariance $\tilde{P}_{k-1}^{(i)}$ of the posterior distribution at time k and hence obtain an approximation $\hat{\Gamma}_k^{\theta_{k-1}^{(i)}}$ of the posterior distribution (update step),
   c. if 1.a or 1.b holds, then for $i = 1, \cdots, N$, compute $\tilde{p}(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})$ using Eq. (2.13), consequently, using (2.18)), obtain an approximation of the normalized weights given

by

$$\tilde{w}_k^{\tilde{\theta}_k^{(i)}} = \frac{\tilde{p}(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})}{\sum_{i=1}^N \tilde{p}(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})},$$

which gives un update to be used in the next step,

2. Resampling: for each $i = 1, \cdots, N$, set $\{\theta_k^{(i)}, B_k^{(i)}, P_k^{(i)}\}$ equal to $\{\tilde{\theta}_k^{(p)}, \tilde{B}_k^{(p)}, \tilde{P}_k^{(p)}\}$ with probability $\tilde{w}_k^{\tilde{\theta}_k^{(p)}}$, where $p \in \{1, \cdots, N\}$.
3. Return to the filtering step.

Note that Algorithm 4.2 can be applied to general state space models (2.1). For our convergence analysis in the next section and in financial applications, we will focus on the special type (3.4), (3.5) of affine state space models. For those models, when $\Sigma \neq 0$, the transition kernel $\pi_k^\theta(\mathrm{d}x \mid x_{k-1})$ resulting (3.4) is not Gaussian, and we need to approximate it by a Gaussian transition kernel in order to apply the Kalman filter.

The approximation is obtained by replacing $\sqrt{x_u^{(1)}}$ in (3.4) with $\sqrt{x_{t_{k-1}}^{(1)}}$, resulting in

$$\begin{aligned}\check{x}_k &= \mathrm{e}^{-A(t_k - t_{k-1})} x_{k-1} + \left(I - \mathrm{e}^{-A(t_k - t_{k-1})}\right)\beta \\ &\quad + \int_{t_{k-1}}^{t_k} \mathrm{e}^{-A(t_k - u)} \left(\Sigma + \tilde{\Sigma}\sqrt{x_{k-1}^{(1)}}\right) \mathrm{d}W_u. \end{aligned} \tag{4.10}$$

Note that given x_{k-1}, the variable $\check{x}_k$ admits a Gaussian transition for $k \in \mathbb{N}^+$. Hence, given the model parameters θ (i.e. A, β, Σ, $\tilde{\Sigma}$), we can compute the approximated transition measure $\hat{\pi}_k^\theta = p(d\check{x}_k \mid x_{k-1}, \theta)$. Recall from Eqs. (2.17), (2.8) and (2.10) that the weights for the parameters space are computed by

$$w_k^{\tilde{\theta}_k^{(i)}} \propto \left((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right). \tag{4.11}$$

In the recursive step of Algorithm 4.2, the measure $\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}$ is not available at time $k-1$, since the parameter evolves from $\theta_{k-1}^{(i)}$ to $\tilde{\theta}_k^{(i)}$ in the jittering step at time k. In order to have a recursive algorithm, the estimate $\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}$ obtained at time $k-1$ is used to approximate the measure $\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}$. Hence, when the model is linear and Gaussian, we obtain the following estimation of the weights for the parameters space

$$\tilde{w}_k^{\tilde{\theta}_k^{(i)}} \propto \left((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}\right). \tag{4.12}$$

When the model is nonlinear or non-Gaussian, then we consider the approximation (4.10) to (3.4). In this case, the transition probability $\pi_k^{\tilde{\theta}_k^{(i)}}$ is approximated by a Gaussian transition probability $\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}$. Therefore, in the non-recursive step, estimation of the weights for the parameters space is given by

$$\hat{w}_k^{\tilde{\theta}_k^{(i)}} \propto \left((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\tilde{\theta}_k^{(i)}}\right), \tag{4.13}$$

In the recursive step, the measure $\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}$ is approximated by $\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}$. Hence, we obtain the following estimation of the weights for the parameters space

$$\hat{w}_k^{\tilde{\theta}_k^{(i)}} \propto \left((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}} \right), \tag{4.14}$$

together with the estimation of the posterior measure of x_k, for $A \in \mathcal{B}(\mathbb{R}^d)$ a Borel set,

$$\hat{\Gamma}_k^{\tilde{\theta}_k^{(i)}}(A) = \frac{\left((\mathbf{1}_{\{x_k \in A\}} l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}} \right)}{\left((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}} \right)}.$$

We will make use of the weights (4.12), (4.13) and (4.14) in our convergence analysis later in Sect. 5.

Remark 4.2 Switching between jittering kernel happens once the condition in the step 1.b of the algorithm is satisfied. Here one may take V_N having diagonal elements of order $N^{-(p+2)/2}$, as mentioned at the beginning of the recursive step. It may however happen in an implementation that the switching does not take place. As will be explained at the beginning of Sect. 5, not switching from kernel 1 to kernel 2 will not prevent convergence of the algorithm. We consider two causes for not switching between kernels: first, the length of the data is too short; and second, the noise level h is extremely big and hence there is not much information contained in the data.

In the first case there are not many data points, and the non-recursive algorithm will anyway quickly produce results, so this case is not problematic from a practical point of view. The second case is not problematic either, because it deals with a really extreme situation. For example, the experiment that we present in Sect. 6.1 on simulated interest rates without noise produces data with on average a magnitude of several basis points (10^{-4}). But the variance of the added noise is 10^{-7}, which is already quite big compared to the interest rate themselves. Even in this case, we observed that after 1164 steps the algorithm switches kernel. Data with 1164 working days (a bit less than 5 years) or more are widely available in financial markets, especially for interest rates. Nevertheless, for Algorithm 4.2 the convergence to the true measure is always assured.

4.2 Kalman Particle filter for models with piece-wise constant parameters

So far in this paper, the model parameters are assumed to be fixed over time. But in many applications, it is more realistic to assume that, at least, some parameters are time-varying, for instance if they are piecewise constant. An offline method that can deal with the estimation of the location of the change-points and of the parameters of the model is studied e.g. in Chib (1998). For an SMC perspective, there are multiple papers that deal with change-point detection problems, see e.g. Chopin (2007), Fearnhead and Liu (2007) and He and Maheu (2010). For the particle filter methods which treat the model parameters as static, the variance of the samples for the parameters decreases with more observed data. Hence the marginal distribution of the model parameters will be increasingly concentrated around certain values. The consequence is that the particle filter algorithm is not able to capture abrupt changes of parameters.

We extend our proposed algorithm for static parameter to adapt to abrupt changes of parameters. To achieve that, we first identify the change points. This step is done by comparing the marginal likelihood between two consecutive steps. Suppose the parameter samples have already converged to the actual value. If at some point the actual parameter value jumps to another value, then the marginal likelihood based on the existing parameter samples are far from optimal. Hence the marginal likelihood at this time point should be significantly smaller than that at the previous time point. On the other hand, if at this time point the actual parameter value does not change, then the marginal likelihood should also be very close to the previous value. So we set a threshold $q < 1$ and if at some point time k, the maximum marginal likelihood for $\theta_k^{(i)} \in D_\theta$, $k \in \mathbb{N}^+$, satisfies

$$\max_{1 \le i \le N} \left\{ p(y_k \mid y_{1:k-1}, \theta_k^{(i)}) \right\} < q \max_{1 \le i \le N} \left\{ p(y_{k-1} \mid y_{1:k-2}, \theta_{k-1}^{(i)}) \right\} ,$$

then we consider the time point k to be the change point of the parameters. Of course, this condition is not sufficient but it is necessary. Notice that the threshold q in our case is user-specific. For statistical methods to choose q, we refer e.g. to Maheu and Gordon (2008).

Another issue here is that the parameter samples may have already (nearly) converged before the change point, hence the variance of these samples is too small to capture the change. This problem can be tackled by adding new samples from the initial parameter distribution to increase the sample variance. But since the variance is increased, the jittering kernel (4.9) does not satisfy Assumption 4.1. Hence the jittering kernel should switch to (4.5). Moreover, since the model parameter changes, the posterior distributions from the previous time point are also not valid anymore, and a new initial value for the mean and the variance of the posterior distribution should also be initialized. So once the change point is determined, one can treat the calibration of the model as a new calibration based on data after this change point.

We introduce the Kalman particle algorithm extended to time-varying parameters. We present the algorithm in full detail, noting that the differences with the previous Algorithm 4.2 are in the two jittering cases in the filtering step.

Algorithm 4.3 (**Kalman particle filter for models with time-varying parameters**)
Initialization:

1. set the number of particles N, a value for the discounting factor a, a switching variance level V_N, a floored variance level V_f, and the threshold parameter q,
2. assume an initial distribution $p(\theta_0)$ for the parameters,
3. sample from the initial distribution to get N particles $\{\theta_0^{(i)}, i = 1, \cdots, N\}$ for the parameters,
4. for each particle $\{\theta_0^{(i)}\}$, assign the same initial mean and covariance value of the posterior distribution $B_0^{(i)}$ and $P_0^{(i)}$ for the Kalman filter update,

Recursion:

1. Filtering: given $\{\theta_{k-1}^{(i)}, i = 1, \cdots, N\}$,
   a. as long as $\operatorname{diag}(\Sigma(\theta_{k-1}))$ is not smaller than V_N (jittering case 1), for each $i = 1, \cdots, N$,
      i. sample new parameters from the kernel (4.5), i.e.,

$$\tilde{\theta}_k^{(i)} \sim \text{TruncNorm}(\tilde{\theta}_k; \mu(\theta_{k-1}^{(i)}), \Sigma(\theta_{k-1}), D_\theta) ,$$

      ii. based on the parameter $\tilde{\theta}_k^{(i)}$, use the Kalman filter to compute the mean and the covariance of the posterior distribution from time 1 to k,

iii. compute the likelihood $p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})$, consequently obtain the normalized weights

$$w_k^{\tilde{\theta}_k^{(i)}} = \frac{p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})}{\sum_{i=1}^N p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})};$$

b. once $\mathrm{diag}(\Sigma(\theta_{l-1})) < V_N$ (jittering case 2), then for $k \geq l+1$ and $i = 1, \cdots, N$, given $\{B_{k-1}^{(i)}, P_{k-1}^{(i)}\}$,

i. sample new parameters from the kernel (4.9), i.e.

$$\tilde{\theta}_k^{(i)} \sim \mathrm{TruncNorm}(\tilde{\theta}_k; \theta_{k-1}^{(i)}, \min\{\max\{\Sigma(\theta_{k-1}), V_f\}, V_N\}, D_\theta),$$

ii. based on the parameters $\tilde{\theta}_k^{(i)}$, $B_{k-1}^{(i)}$ and $P_{k-1}^{(i)}$, use the Kalman filter to compute the mean $\tilde{B}_k^{(i)}$ and the covariance $\tilde{P}_k^{(i)}$ of the approximated posterior distribution at time k,

iii. compute an approximation $\tilde{p}(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})$ of the likelihood $p(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})$. If

$$\max_{1\leq i\leq N}\{\tilde{p}(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})\} < q \max_{1\leq i\leq N}\{\tilde{p}(y_{k-1} \mid y_{1:k-2}, \tilde{\theta}_{k-1}^{(i)})\},$$

then go to the Initialization step of the algorithm to initialize the algorithm using the data after time k. Otherwise compute the normalized weights

$$\tilde{w}_k^{\tilde{\theta}_k^{(i)}} = \frac{\tilde{p}(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})}{\sum_{i=1}^N \tilde{p}(y_k \mid y_{1:k-1}, \tilde{\theta}_k^{(i)})}.$$

2. Resampling: for each $i = 1, \cdots, N$, set $\{\theta_k^{(i)}, B_k^{(i)}, P_k^{(i)}\} = \{\tilde{\theta}_k^{(p)}, \tilde{B}_k^{(p)}, \tilde{P}_k^{(p)}\}$, with probability $w_k^{\tilde{\theta}_k^{(p)}}$, where $p \in \{1, \cdots, N\}$.

In Sect. 6 we will present a numerical study where Algorithm 4.3 is seen to be capable of quickly tracking a sudden parameter change. Similar to the example in Crisan and Míguez (2018), Sect. 6, we content ourselves with this empirical tracking behaviour and refrain from presenting a consistency result in a large sample setting. For a fixed number of change points though, convergence results do exist, see e.g. Bai and Perron (1998) on multiple structural changes in linear regression. But, as stated above, an asymptotic analysis of this behaviour is besides the purpose of the present section.

5 Convergence analysis

This section is devoted to showing a convergence result, Theorem 5.9 below, for the Kalman particle Algorithm 4.2 introduced in Sect. 4.1.2. The standing assumption is that we have observations $y_{1:k}$ generated by (3.4) and (3.5). The parameters are $A, \beta, \Sigma, \tilde{\Sigma}$, collectively denoted θ.

In Algorithm 4.2, the (conditional) measure

$$\mu_k(\mathrm{d}\theta_k) = p(\mathrm{d}\theta_k \mid y_{1:k}), \tag{5.1}$$

is estimated for $k = 1, \cdots, K$. From now on, for convenience, we assume the algorithm starts at time 1. At each time, the algorithm has three main steps: jittering, update and resampling.

Define the maximum step size $\Delta = \sup_{k=1,\cdots,K}(t_k - t_{k-1})$. Let the parameter $\theta \in D_\theta$. At time k, suppose the estimated measure of the last time μ_{k-1}^N is available and

$$\mu_{k-1}^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_{k-1}^{(i)}}(\mathrm{d}\theta)\,.$$

In the jittering step as described in Algorithm 4.2, new samples $\tilde{\theta}_k^{(i)}$ are sampled from the kernel functions. The resulting measure $\tilde{\mu}_k^N$ is then defined by

$$\tilde{\mu}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{\theta}_k^{(i)}}(\mathrm{d}\theta)\,. \tag{5.2}$$

In this step, no extra information is used to refine the estimates on the parameters. Hence the aim is to prove that the measure $\tilde{\mu}_k^N$ converges to the measure μ_{k-1}^N in some sense when Δ goes to zero and the number of samples N goes to infinity.

In the update step as described in Algorithm 4.2 of Sect. 4.1.2, there are four cases to analyze, combinations of Gaussian-linear or non-Gaussian/non-linear models and recursive or non-recursive parts of the algorithm. When the model is linear and Gaussian, and we consider the non-recursive step of the algorithm, then the normalized weights $w_k^{\tilde{\theta}_k^{(i)}}$ are computed exactly. When we consider the recursive step of the algorithm for a Gaussian linear model, the normalized weights $w_k^{\tilde{\theta}_k^{(i)}}$ are estimated by $\tilde{w}_k^{\tilde{\theta}_k^{(i)}}$, see (4.12). For the non-Gaussian and non-linear model, we consider the approximation (4.10) to (3.4), the normalized weights are estimated by (4.13) and (4.14) in the non-recursive step and recursive step, respectively. Since the convergence of the weights in this latter case, the recursive step, *implies the convergence of the weights in the first three cases*, we will only consider the model (4.10) and the recursive step of the algorithm. In this case, we obtain a new estimated measure

$$\hat{\mu}_k^N = \sum_{i=1}^N \hat{w}_k^{\tilde{\theta}_k^{(i)}} \delta_{\tilde{\theta}_k^i}(\mathrm{d}\theta) \tag{5.3}$$

and the convergence analysis of $\hat{\mu}_k^N$ depends on that of the estimated weights $\hat{w}_k^{\tilde{\theta}_k^{(i)}}$. The aim is to prove that $\hat{\mu}_k^N$ converges to μ_k in some sense when Δ goes to 0 and N goes to infinity.

In the resampling step, the new particles $\{\theta_k^{(i)},\ i = 1, \cdots, N\}$ are sampled from the empirical distribution $\sum_{i=1}^N \hat{w}_k^{\tilde{\theta}_k^{(i)}} \delta_{\tilde{\theta}_k^{(i)}}$ and we need to prove that the measure

$$\mu_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_k^{(i)}}(\mathrm{d}\theta)\,, \tag{5.4}$$

converges to μ_k when Δ goes to zero and N goes to infinity.

For our convergence analysis we will need besides Assumption 4.1, the following three assumptions. They are of a type that is common when dealing with parameter estimation using particle filters.

Assumption 5.1 The posterior measure Γ_k^θ, $k = 1, \cdots, N$, $\theta \in D_\theta$, is Lipschitz in the parameter θ, i.e. for any bounded continuous function f,

$$\left|(f, \Gamma_k^\theta) - (f, \Gamma_k^{\theta'})\right| \le e_{3,k}\, \|f\|_\infty \left\|\theta - \theta'\right\|\,, \tag{5.5}$$

for some positive constant $e_{3,k}$ independent of f .

Assumption 5.2 Let $(x_k)_{k\in\mathbb{N}^+}$ be as in (3.4). To emphasize that the dependence of the distribution of x_k depends on θ, we write $(x_k^\theta)_{k\in\mathbb{N}^+}$. It is assumed that there exist a constant $M > 0$ such that

$$\sup_{\theta\in D_\theta,\, k\in\mathbb{N}^+} \mathbb{E}\left[\left|x_k^\theta\right|\right] \leq M.$$

Assumption 5.3 For any fixed observation sequence $y_{1:k}$, the likelihood $\{l_{y_t}^\theta(x),\ t = 1,\cdots,k,\ \theta\in D_\theta\}$ satisfies

1. $\left\|l_{y_t}\right\|_\infty := \sup_{\theta\in D_\theta}\left\|l_{y_t}^\theta\right\|_\infty < \infty$,
2. $\inf_{\theta\in D_\theta} l_{y_t}^\theta > 0$.

Remark 5.1 Assumptions 5.2 and 5.3 hold when a judicious choice of D_θ is made. For instance, the inequality in Assumption 5.2 involves x_k^θ for all k. This leads us to assume that the matrix $-A$ is stable, all eigenvalues of A have positive real part, or in the diagonal case, all diagonal elements of A are positive. To have also Assumption 5.3 satisfied, one has to impose further restrictions, the elements of A come from a bounded set and its eigenvalues are bounded away from zero (then also A^{-1} belongs to a bounded set), whereas these conditions are also imposed on Σ and $\tilde{\Sigma}$. Under these specified conditions, Assumptions 5.2 and 5.3 hold true.

The Inequality (4.6) in Assumption 4.1 is used for the convergence analysis of the jittering step, the Inequality (4.7) in Assumption 4.1, together with Assumptions 5.1 and 5.2 are specially used for the analysis for the update step and Assumption 5.3 is used for the analysis for both update and resampling steps. Given the convergence of μ_{k-1}^N to μ_{k-1} in an L^p-sense when Δ goes to zero and N goes to infinity, we present the convergence of the measures $\tilde{\mu}_k^N$, $\hat{\mu}_k^N$ and μ_k^N respectively in the three Lemmas 5.4, 5.7, 5.8 in the subsections below. In Sect. 5.4, we study the convergence of Algorithm 4.2 presented in Sect. 4.1.2 by induction using these three lemmas. The main result there is Theorem 5.9. Our analysis in inspired by the results in Crisan and Míguez (2018) for the nested particle filter and follows a similar pattern. Note however the crucial differences between our Algorithm 4.2 and their nested particle filter. We have to deal with only one layer with a particle filter (instead of two such layers), as we use the Kalman filter in the other layer, but we also pay attention to the time discretization of the continuous processes.

5.1 Jittering

In the jittering step 1.b.i for low values of $\Sigma(\theta_{k-1})$ the new parameter particles $\tilde{\theta}_k^{(i)}$ are sampled from a kernel function $\kappa_k^N(\mathrm{d}\theta \mid \theta_{k-1}^{(i)})$, $i = 1,\cdots,N$. The following lemma shows that the error due to the jittering step vanishes. This lemma can be seen as the analog of Lemma 3 in Crisan and Míguez (2018), but with the term $\frac{1}{\sqrt{M}}$ there (which plays no role in our analysis) replaced with $\sqrt{\Delta}$ as we treat the influence of time discretization. It can be proven in the same way and we present it here for the sake of completeness and in a form that suits our purposes.

Lemma 5.4 *Let f be a bounded function and suppose Assumption 4.1 holds. If*

$$\left\|(f,\mu_{k-1}^N) - (f,\mu_{k-1})\right\|_p \leq \frac{c_{1,k-1}\,\|f\|_\infty}{\sqrt{N}} + d_{1,k-1}\,\|f\|_\infty\sqrt{\Delta}, \tag{5.6}$$

for some constants $c_{1,k-1}$ and $d_{1,k-1}$ which are independent of f, N and Δ, then there exist constants $\tilde{c}_{1,k}$ and $\tilde{d}_{1,t}$ which are independent of f, N and Δ, such that

$$\left\|(f,\tilde{\mu}_k^N)-(f,\mu_{k-1})\right\|_p \leq \frac{\tilde{c}_{1,k}\,\|f\|_\infty}{\sqrt{N}}+\tilde{d}_{1,k}\,\|f\|_\infty\sqrt{\Delta}\,. \tag{5.7}$$

5.2 Update

To prove the convergence in the update step 1.c, we first need to prove that the error introduced from the approximation of the weights $w_k^{\tilde{\theta}_k^{(i)}}$ by the weights $\hat{w}_k^{\tilde{\theta}_k^{(i)}}$, can be bounded by a desired quantity as in (5.7). The following lemma is a core result in our convergence analysis, in the proof of it we exploit the affine nature of the state process.

Lemma 5.5 *Let the observation sequence $y_{1:k}$ be fixed. Suppose function f is bounded and continuous and Assumptions 4.1,5.1, and 5.2 hold. If*

$$\sup_{1\leq i\leq N}\left|(f,\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-(f,\Gamma_{k-1}^{\theta_{k-1}^{(i)}})\right| \leq \frac{c_{2,k-1}\,\|f\|_\infty}{\sqrt{N}}+d_{2,k-1}\,\|f\|_\infty\sqrt{\Delta}\,, \tag{5.8}$$

for some constants $c_{2,k-1}$ and $d_{2,k-1}$ which are independent of f, N and Δ, then there exist constants $\tilde{c}_{2,k}$ and $\tilde{d}_{2,k}$ which are independent of f, N and Δ such that

$$\sup_{1\leq i\leq N}\left|((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-((f,\pi_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right| \leq \frac{\tilde{c}_{2,k}\,\|f\|_\infty}{\sqrt{N}}+\tilde{d}_{2,k}\,\|f\|_\infty\sqrt{\Delta}\,. \tag{5.9}$$

To prove Lemma 5.5, we exploit the structure (3.4) of the state process. Besides, we need the following auxiliary result. The proof of it follows the same lines as the proof of Lemma 4 in Crisan and Míguez (2018), but note again the $\frac{1}{\sqrt{M}}$ term is replaced by $\sqrt{\Delta}$.

Lemma 5.6 *Suppose the function f is bounded and continuous. Moreover, suppose Assumptions 4.1 and 5.1 and Inequality (5.8) hold. Then there exist some constants $\tilde{c}_{2,k-1}$ and $\tilde{d}_{2,k-1}$ which are independent of f, N, Δ and of all θ such that*

$$\sup_{1\leq i\leq N}\left|(f,\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-(f,\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right| \leq \frac{\tilde{c}_{2,k-1}\,\|f\|_\infty}{\sqrt{N}}+\tilde{d}_{2,k}\,\|f\|_\infty\sqrt{\Delta}\,. \tag{5.10}$$

Now we are ready to prove Lemma 5.5.

Proof Using the triangle inequality, one obtains

$$\begin{aligned}
&\sup_{1\leq i\leq N}\left|((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-((f,\pi_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right|\\
&\quad\leq \sup_{1\leq i\leq N}\left|((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right|\\
&\qquad+\sup_{1\leq i\leq N}\left|((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})-((f,\pi_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right|\,.
\end{aligned} \tag{5.11}$$

Note that $\sup_{\theta\in D_\theta}(f,\hat{\pi}_k^\theta)$ is bounded by $\|f\|_\infty$. Hence, using Inequality (5.10), we get for the first term on the right hand side of (5.11)

$$
\begin{aligned}
&\sup_{1\le i\le N}\left|((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right|\\
&\quad\le\sup_{\theta\in D_\theta}\sup_{1\le i\le N}\left|((f,\hat{\pi}_k^{\theta}),\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-((f,\hat{\pi}_k^{\theta}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right|\\
&\quad\le\frac{\tilde{c}_{2,k-1}\,\|f\|_\infty}{\sqrt{N}}+\tilde{d}_{2,k}\,\|f\|_\infty\sqrt{\Delta}\,. \qquad (5.12)
\end{aligned}
$$

Recall $(x_k)_{k\in\mathbb{N}}$ and $(\check{x}_k)_{k\in\mathbb{N}}$ respectively from (3.4) and (4.10). For $\epsilon, M_1 > 0$, define the sets

$$
\begin{aligned}
A_{\epsilon,i}&=\left\{\left|\check{x}_k^{\tilde{\theta}_k^{(i)}}-x_k^{\tilde{\theta}_k^{(i)}}\right|<\epsilon\right\},\\
B_{M_1,i}&=\left\{\left|\check{x}_k^{\tilde{\theta}_k^{(i)}}\right|\le M_1,\ \left|x_k^{\tilde{\theta}_k^{(i)}}\right|\le M_1\right\}.
\end{aligned}
$$

Note that $(f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}})$ can be seen as the (conditional) expectation of $f(.)$ taken under the measure $\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}$. Stated otherwise, we can see it as the expectation of $f(\check{x}_k^{\tilde{\theta}_k^{(i)}})$. Likewise, we can see $(f,\pi_k^{\tilde{\theta}_k^{(i)}})$ as the (conditional) expectation of $f(x_k^{\tilde{\theta}_k^{(i)}})$. Below we use the notations $\mathbb{E}f(\check{x}_k^{\tilde{\theta}_k^{(i)}})$ and $\mathbb{E}f(x_k^{\tilde{\theta}_k^{(i)}})$ for these expectations. With these interpretations, the second term on the right hand side of (5.11) yields

$$
\begin{aligned}
&\sup_{1\le i\le N}\left|\left((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)-\left((f,\pi_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)\right|\\
&\quad=\sup_{1\le i\le N}\left|\left((f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}})-(f,\pi_k^{\tilde{\theta}_k^{(i)}}),\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)\right|\\
&\quad\le\sup_{1\le i\le N}\left(\left|(f,\hat{\pi}_k^{\tilde{\theta}_k^{(i)}})-(f,\pi_k^{\tilde{\theta}_k^{(i)}})\right|,\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)\\
&\quad=\sup_{1\le i\le N}\left(\left|\mathbb{E}f(\check{x}_k^{\tilde{\theta}_k^{(i)}})-\mathbb{E}f(x_k^{\tilde{\theta}_k^{(i)}})\right|,\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)\\
&\quad\le\sup_{1\le i\le N}\left(\mathbb{E}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}})-f(x_k^{\tilde{\theta}_k^{(i)}})\right|,\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)\\
&\quad=\sup_{1\le i\le N}\left(\mathbb{E}\mathbf{1}_{A_{\epsilon,i}}\mathbf{1}_{B_{M_1,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}})-f(x_k^{\tilde{\theta}_k^{(i)}})\right|,\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)\\
&\qquad+\sup_{1\le i\le N}\left(\mathbb{E}\mathbf{1}_{A_{\epsilon,i}}\mathbf{1}_{B^c_{M_1,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}})-f(x_k^{\tilde{\theta}_k^{(i)}})\right|,\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right)\\
&\qquad+\sup_{1\le i\le N}\left(\mathbb{E}\mathbf{1}_{A^c_{\epsilon,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}})-f(x_k^{\tilde{\theta}_k^{(i)}})\right|,\Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right). \qquad (5.13)
\end{aligned}
$$

We need to find an upper bound for the three terms on the right hand side of (5.13).

Consider the first term on the right hand side of (5.13) and note that the random variables $\check{x}_k^{\tilde{\theta}_k^{(i)}}$ and $x_k^{\tilde{\theta}_k^{(i)}}$ restricted to B_{M_1} take values in a compact set. Hence the continuous function f is also uniformly continuous on that set. Then there exists an $\epsilon > 0$, such that $|f(x) - f(y)| \le \sqrt{\Delta}$, for all $|x - y| < \epsilon$. From now on we assume that we use this ϵ, and we obtain for the first term on the right hand side of (5.13), using the definition of $A_{\epsilon,i}$,

$$\sup_{1\le i\le N} \mathbb{E}\mathbf{1}_{A_{\epsilon,i}}\mathbf{1}_{B_{M_1,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}}) - f(x_k^{\tilde{\theta}_k^{(i)}})\right| \le \sqrt{\Delta}\,,$$

which implies

$$\sup_{1\le i\le N}\left(\mathbb{E}\mathbf{1}_{A_{\epsilon,i}}\mathbf{1}_{B_{M_1,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}}) - f(x_k^{\tilde{\theta}_k^{(i)}})\right|, \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right) \le \sqrt{\Delta}\,. \tag{5.14}$$

Next we consider the second term on the right hand side of (5.13). Set $M_1 = M/\sqrt{\Delta} + \epsilon$, where M is as defined in Assumption 5.2. On the set $A_{\epsilon,i}$, we have $\left|\check{x}_k^{\tilde{\theta}_k^{(i)}}\right| - \epsilon \le \left|x_k^{\tilde{\theta}_k^{(i)}}\right| \le \left|\check{x}_k^{\tilde{\theta}_k^{(i)}}\right| + \epsilon$, which implies $\mathbb{P}\left(A_{\epsilon,i}, \left|\check{x}_k^{\tilde{\theta}_k^{(i)}}\right| > M_1\right) \le \mathbb{P}\left(\left|x_k^{\tilde{\theta}_k^{(i)}}\right| > M_1 - \epsilon\right)$. Hence using Assumption 5.2 and the Markov inequality, we obtain

$$\begin{aligned}
&\sup_{1\le i\le N} \mathbb{E}\mathbf{1}_{A_{\epsilon,i}}\mathbf{1}_{B^c_{M_1,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}}) - f(x_k^{\tilde{\theta}_k^{(i)}})\right| \\
&\le 2\,\|f\|_\infty \sup_{1\le i\le N}\left(\mathbb{E}\mathbf{1}_{A_{\epsilon,i}}\mathbf{1}_{B^c_{M_1,i}}\right) \\
&\le 2\,\|f\|_\infty \sup_{1\le i\le N}\left[\mathbb{P}\left(A_{\epsilon,i}, \left|\check{x}_k^{\tilde{\theta}_k^{(i)}}\right| > M_1\right) + \mathbb{P}\left(A_{\epsilon,i}, \left|x_k^{\tilde{\theta}_k^{(i)}}\right| > M_1\right)\right] \\
&\le 2\,\|f\|_\infty \sup_{1\le i\le N}\left[\mathbb{P}\left(\left|x_k^{\tilde{\theta}_k^{(i)}}\right| > M_1 - \epsilon\right) + \mathbb{P}\left(\left|x_k^{\tilde{\theta}_k^{(i)}}\right| > M_1\right)\right] \\
&\le 4\,\|f\|_\infty \sup_{1\le i\le N}\mathbb{P}\left(\left|x_k^{\tilde{\theta}_k^{(i)}}\right| > M_1 - \epsilon\right) \\
&\le 4\,\|f\|_\infty \frac{\sup_{1\le i\le N}\mathbb{E}\left|x_k^{\tilde{\theta}_k^{(i)}}\right|}{M_1 - \epsilon} \\
&\le 4\,\|f\|_\infty \sqrt{\Delta}\,.
\end{aligned} \tag{5.15}$$

Substituting (5.15) into the second term of the right hand side of (5.13), we obtain

$$\begin{aligned}
&\sup_{1\le i\le N}\left(\mathbb{E}\mathbf{1}_{A_{\epsilon,i}}\mathbf{1}_{B^c_{M_1,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}}) - f(x_k^{\tilde{\theta}_k^{(i)}})\right|, \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right) \\
&\le \sup_{1\le i\le N}\left(4\,\|f\|_\infty \sqrt{\Delta}, \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right) \\
&\le 4\,\|f\|_\infty \sqrt{\Delta}\,.
\end{aligned} \tag{5.16}$$

For the last term on the right hand side of (5.13), we apply again the Markov inequality, to obtain

$$\begin{aligned}
\mathbb{E}\mathbf{1}_{A^c_{\epsilon,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}}) - f(x_k^{\tilde{\theta}_k^{(i)}})\right| &\le 2\,\|f\|_\infty \mathbb{P}(A^c_{\epsilon,i}) \\
&\le 2\,\|f\|_\infty \frac{\mathbb{E}\left|\check{x}_k^{\tilde{\theta}_k^{(i)}} - x_k^{\tilde{\theta}_k^{(i)}}\right|}{\epsilon}\,.
\end{aligned}$$

We denote the i-th component of an $\mathbb{R}^d$-valued process $(x_k)_{k\in\mathbb{N}}$ by $(x_k^{(i)})_{k\in\mathbb{N}}$. Furthermore, we temporarily suppress the dependence on $\tilde{\theta}_k^{(i)}$ in the notation. Given x_{k-1}, according to Eqs. (3.4) and (4.10), we obtain

$$\begin{aligned}\mathbb{E}\left|\check{x}_k^{(i)} - x_k^{(i)}\right| &= \mathrm{e}^{-\alpha_i(t_k-t_{k-1})}\mathbb{E}\left|\int_{t_{k-1}}^{t_k} \mathrm{e}^{\alpha_i u}\sum_{j=1}^d \Sigma_{ij}\left(\sqrt{x_u^{(1)}} - \sqrt{x_{k-1}^{(1)}}\right)\mathrm{d}W_u^{(j)}\right| \\ &\le \mathbb{E}\left|\int_{t_{k-1}}^{t_k} \mathrm{e}^{\alpha_i u}\sum_{j=1}^d \Sigma_{ij}\left(\sqrt{x_u^{(1)}} - \sqrt{x_{k-1}^{(1)}}\right)\mathrm{d}W_u^{(j)}\right| .\end{aligned}$$

Define $\Sigma^{(i)} = \sqrt{\sum_{j=1}^d (\Sigma_{ij})^2}$. Using the Burkholder-Davis-Gundy inequality (Karatzas and Shreve 1998, Theorem 3.28), we know there exists a constant C which does not depend on $(x_t)_{t\ge 0}$ such that, for every $i = 1, \cdots, d$,

$$\begin{aligned}&\mathbb{E}\left|\int_{t_{k-1}}^{t_k} \mathrm{e}^{\alpha_i u}\sum_{j=1}^d \Sigma_{ij}\left(\sqrt{x_u^{(1)}} - \sqrt{x_{k-1}^{(1)}}\right)\mathrm{d}W_u\right| \\ &\le C\Sigma^{(i)}\mathbb{E}\left(\int_{t_{k-1}}^{t_k} \mathrm{e}^{2\alpha_i u}\left(\sqrt{x_u^{(1)}} - \sqrt{x_{k-1}^{(1)}}\right)^2 \mathrm{d}u\right)^{\frac{1}{2}} \\ &\le C\Sigma^{(i)}\mathbb{E}\left(\int_{t_{k-1}}^{t_k} \mathrm{e}^{2\alpha_i u}\left(x_u^{(1)} + x_{k-1}^{(1)}\right)\mathrm{d}u\right)^{\frac{1}{2}} .\end{aligned}$$

Using Jensen's inequality and Fubini's theorem, we get for the latter expectation

$$\begin{aligned}&\mathbb{E}\left(\int_{t_{k-1}}^{t_k} \mathrm{e}^{2\alpha_i u}\left(x_u^{(1)} + x_{k-1}^{(1)}\right)\mathrm{d}u\right)^{\frac{1}{2}} \\ &\le \left(\mathbb{E}\int_{t_{k-1}}^{t_k} \mathrm{e}^{2\alpha_i u}\left(x_u^{(1)} + x_{k-1}^{(1)}\right)\mathrm{d}u\right)^{\frac{1}{2}} \\ &= \left(\int_{t_{k-1}}^{t_k} \mathrm{e}^{2\alpha_i u}\mathbb{E}\left(x_u^{(1)} + x_{k-1}^{(1)}\right)\mathrm{d}u\right)^{\frac{1}{2}} \\ &= \left(\int_{t_{k-1}}^{t_k} \mathrm{e}^{2\alpha_i u}\left[(1+\mathrm{e}^{-\alpha_1(u-t_{k-1})})x_{k-1}^{(1)} + (1-e^{-\alpha_1(u-t_{k-1})})\beta_1\right]\mathrm{d}u\right)^{\frac{1}{2}} \\ &= \left(\frac{x_{k-1}^{(1)}+\beta_1}{2\alpha_i}\left(\mathrm{e}^{2\alpha_i\Delta}-1\right)e^{2\alpha_i t_{k-1}}\right. \\ &\quad \left.+\frac{x_{k-1}^{(1)}-\beta_1}{2\alpha_i-\alpha_1}\left(\mathrm{e}^{(2\alpha_i-\alpha_1)\Delta}-1\right)e^{2\alpha_i t_{k-1}}\right)^{\frac{1}{2}} .\end{aligned}$$

Note that $\mathrm{e}^x - 1 = O(x)$ if $x \to 0$. Since the parameters are assumed to have a compact domain, we conclude there exists a constant C_1 which is independent of the parameters and such that

$$\mathbb{E}\left|\check{x}_k - x_k\right| \le C_1 x_{k-1}^{(1)}\sqrt{\Delta} .$$

Hence, returning to previously used notation,

$$\mathbb{E}\mathbf{1}^{c}_{A_{\epsilon,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}}) - f(x_k^{\tilde{\theta}_k^{(i)}})\right| \le \frac{C_1}{\epsilon} x_{k-1}^{(1),\tilde{\theta}_k^{(i)}}\sqrt{\Delta}\,,$$

which implies

$$\begin{aligned}\sup_{1\le i\le N}\left(\mathbb{E}\mathbf{1}_{A^{c}_{\epsilon,i}}\left|f(\check{x}_k^{\tilde{\theta}_k^{(i)}}) - f(x_k^{\tilde{\theta}_k^{(i)}})\right|, \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}\right) &\le \frac{C_1}{\epsilon}\sqrt{\Delta}\sup_{1\le i\le N}\mathbb{E}x_{k-1}^{(1),\tilde{\theta}_k^{(i)}}\\ &\le \frac{C_1 M}{\epsilon}\sqrt{\Delta}\,. \qquad (5.17)\end{aligned}$$

Combining (5.16),(5.14) and (5.17) together with (5.12), we prove the statement of the lemma. □

Lemma 5.5 shows that the approximation errors of $\hat{\pi}_k^{\tilde{\theta}_k^{(i)}}$ and $\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}$ can be controlled in an appropriate manner, guaranteeing $\Sigma(\theta_{k-1})$ below a threshold value V_N. This allows us to run the outer layer in Algorithm 4.2 recursively, see step 1.b. Adding Assumption 5.3, we present in the following lemma the convergence of $\hat{\mu}_k^N$.

Lemma 5.7 *Let the observation sequence $y_{1:k}$ be fixed and Assumptions 4.1, 5.1, 5.2 and 5.3 hold. Then for any bounded and continuous function f, if (5.6) and*

$$\sup_{1\le i\le N}\left|(f, \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}) - (f, \Gamma_{k-1}^{\theta_{k-1}^{(i)}})\right| \le \frac{c_{2,k-1}\left\|f\right\|_\infty}{\sqrt{N}} + d_{2,k-1}\left\|f\right\|_\infty\sqrt{\Delta}\,,$$

hold for some constants $c_{2,k-1}$ and $d_{2,k-1}$ which are independent of f, N and Δ, then there exist constants $\hat{c}_{1,k}$, $\hat{d}_{1,k}$, $\tilde{c}_{2,k}$ and $\tilde{d}_{2,k}$ which are independent of f, N and Δ such that

$$\begin{aligned}\left\|(f, \hat{\mu}_k^N) - (f, \mu_k)\right\|_p &\le \frac{\hat{c}_{1,k}\left\|f\right\|_\infty}{\sqrt{N}} + \hat{d}_{1,k}\left\|f\right\|_\infty\sqrt{\Delta}\,,\\ \sup_{1\le i\le N}\left|(f, \hat{\Gamma}_k^{\tilde{\theta}_k^{(i)}}) - (f, \Gamma_k^{\tilde{\theta}_k^{(i)}})\right| &\le \frac{\tilde{c}_{2,k}\left\|f\right\|_\infty}{\sqrt{N}} + \tilde{d}_{2,k}\left\|f\right\|_\infty\sqrt{\Delta}\,.\end{aligned}$$

Proof First, we prove the convergence of the estimated normalized weights $\hat{w}_k^{\tilde{\theta}_k^{(i)}}$, which are used to prove the convergence of the measures $\hat{\mu}_k^N$. Denote the unnormalized weights by $\hat{v}_k^{\tilde{\theta}_k^{(i)}} = ((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})$, see (4.14), and $v_k^{\tilde{\theta}_k^{(i)}} = ((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})$, see (4.11). Note that $\sup_{\theta\in D_\theta}(l_{y_k}^{\theta}, \pi_k^{\theta})$ is bounded by $\left\|l_{y_k}\right\|_\infty$. Using Lemma 5.5, noting that $\left\|l_{y_t}\right\|_\infty$ is some finite number under Assumption 5.3, we have

$$\begin{aligned}\sup_{1\le i\le N}\left|\hat{v}_k^{\tilde{\theta}_k^{(i)}} - v_k^{\tilde{\theta}_k^{(i)}}\right| &= \sup_{1\le i\le N}\left|((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}) - ((l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right|\\ &\le \sup_{\theta\in D_\theta}\sup_{1\le i\le N}\left|((l_{y_k}^{\theta}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}) - ((l_{y_k}^{\theta}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right|\\ &\le \frac{\tilde{c}_{2,k}}{\sqrt{N}} + \tilde{d}_{2,k}\sqrt{\Delta}\,, \qquad (5.18)\end{aligned}$$

where $\tilde{c}_{2,k}$ and $\tilde{d}_{2,k}$ are constants which are independent of N, Δ and θ. By Assumption 5.3, we obtain

$$\inf_{\theta\in D_\theta} \hat{v}_k^\theta, \inf_{\theta\in D_\theta} v_k^\theta > 0\,,$$
$$\sup_{\theta\in D_\theta} \hat{v}_k^\theta, \sup_{\theta\in D_\theta} v_k^\theta < \infty\,.$$

Hence for the normalized weights, it follows that

$$\begin{aligned}
\sup_{1\le i\le N}\left|\hat{w}_k^{\tilde{\theta}_k^{(i)}} - w_k^{\tilde{\theta}_k^{(i)}}\right| &= \sup_{1\le i\le N}\left|\frac{\hat{v}_k^{\tilde{\theta}_k^{(i)}}}{\sum_{i=1}^N \hat{v}_k^{\tilde{\theta}_k^{(i)}}} - \frac{v_k^{\tilde{\theta}_k^{(i)}}}{\sum_{i=1}^N v_k^{\tilde{\theta}_k^{(i)}}}\right| \\
&\le \sup_{1\le i\le N}\left|\frac{\hat{v}_k^{\tilde{\theta}_k^{(i)}}}{\sum_{i=1}^N \hat{v}_k^{\tilde{\theta}_k^{(i)}}} - \frac{\hat{v}_k^{\tilde{\theta}_k^{(i)}}}{\sum_{i=1}^N v_k^{\tilde{\theta}_k^{(i)}}}\right| \\
&\quad + \sup_{1\le i\le N}\left|\frac{\hat{v}_k^{\tilde{\theta}_k^{(i)}}}{\sum_{i=1}^N v_k^{\tilde{\theta}_k^{(i)}}} - \frac{v_k^{\tilde{\theta}_k^{(i)}}}{\sum_{i=1}^N v_k^{\tilde{\theta}_k^{(i)}}}\right| \\
&\le \sup_{1\le i\le N}\frac{\hat{v}_k^{\tilde{\theta}_k^{(i)}}}{(\sum_{i=1}^N v_k^{\tilde{\theta}_k^{(i)}})(\sum_{i=1}^N \hat{v}_k^{\tilde{\theta}_k^{(i)}})}\sum_{i=1}^N\left|v_k^{\tilde{\theta}_k^{(i)}} - \hat{v}_k^{\tilde{\theta}_k^{(i)}}\right| \\
&\quad + \sup_{1\le i\le N}\frac{1}{\sum_{i=1}^N \hat{v}_k^{\tilde{\theta}_k^{(i)}}}\left|\hat{v}_k^{\tilde{\theta}_k^{(i)}} - v_k^{\tilde{\theta}_k^{(i)}}\right|\,. \qquad (5.19)
\end{aligned}$$

Observe that $\frac{\hat{v}_k^{\tilde{\theta}_k^{(i)}}}{\sum_{i=1}^N v_k^{\tilde{\theta}_k^{(i)}}} \le 1$ and that $\sum_{i=1}^N \hat{v}_k^{\tilde{\theta}_k^{(i)}}$ is bounded from below by a constant times N. Substituting Inequality (5.18) into Inequality (5.19), one obtains that there exist constants $\hat{c}_{2,k}$ and $\hat{d}_{2,k}$ such that

$$\sup_{1\le i\le N}\left|\hat{w}_k^{\tilde{\theta}_k^{(i)}} - w_k^{\tilde{\theta}_k^{(i)}}\right| \le \frac{\hat{c}_{2,k}}{\sqrt{N}} + \hat{d}_{2,k}\sqrt{\Delta}\,. \qquad (5.20)$$

Next we study the convergence of the measures $\hat{\mu}_k^N$. For simplification in the notation, we write $\hat{w}_k = \hat{w}_k^{\tilde{\theta}_k^{(i)}}$. Recalling that $w_k = \frac{p(y_k|y_{1:k-1},\theta)}{\int p(y_k|y_{1:k-1},\theta)d\theta}$ (note that $\|w_k\|_\infty < \infty$ by Assumption 5.3)), we get from Bayes' rule

$$\begin{aligned}
(f,\mu_k) &= \int f(\theta)p(\theta \mid y_{1:k})d\theta \\
&= \int f(\theta)\frac{p(y_k \mid y_{1:k-1},\theta)p(\theta \mid y_{1:k-1})}{\int p(y_k \mid y_{1:k-1},\theta)p(\theta \mid y_{1:k-1})d\theta}d\theta \\
&= \frac{(f w_k,\mu_{k-1})}{(w_k,\mu_{k-1})}\,,
\end{aligned}$$

and from (5.2) and (5.3) we get

$$(f,\hat{\mu}_k^N) = \frac{(f\hat{w}_k,\tilde{\mu}_k^N)}{(\hat{w}_k,\tilde{\mu}_k^N)}\,.$$

Since w_k and $(w_k,\mu_{k-1}) \ge 0$, using Assumption 5.3 and the triangle inequality, we get

$$
\begin{aligned}
&\left\|(f,\hat{\mu}_k^N)-(f,\mu_k)\right\|_p\\
&=\left\|\frac{(f\hat{w}_k,\tilde{\mu}_k^N)}{(\hat{w}_k,\tilde{\mu}_k^N)}-\frac{(fw_k,\mu_{k-1})}{(w_k,\mu_{k-1})}\right\|_p\\
&=\frac{1}{(w_k,\mu_{k-1})}\left\|\frac{(f\hat{w}_k,\tilde{\mu}_k^N)}{(\hat{w}_k,\tilde{\mu}_k^N)}(w_k,\mu_{k-1})-(fw_k,\mu_{k-1})\right\|_p\\
&\le\frac{1}{(w_k,\mu_{k-1})}\left(\left\|\frac{(f\hat{w}_k,\tilde{\mu}_k^N)}{(\hat{w}_k,\tilde{\mu}_k^N)}(w_k,\mu_{k-1})-(f\hat{w}_k,\tilde{\mu}_k^N)\right\|_p\right.\\
&\quad\left.+\left\|(f\hat{w}_k,\tilde{\mu}_k^N)-(fw_k,\mu_{k-1})\right\|_p\right)\\
&=\frac{1}{(w_k,\mu_{k-1})}\left(\left\|\frac{(f\hat{w}_k,\tilde{\mu}_k^N)[(w_k,\mu_{k-1})-(\hat{w}_k,\tilde{\mu}_k^N)]}{(\hat{w}_k,\tilde{\mu}_k^N)}\right\|_p\right.\\
&\quad\left.+\left\|(f\hat{w}_k,\tilde{\mu}_k^N)-(fw_k,\mu_{k-1})\right\|_p\right)\\
&\le\frac{1}{(w_k,\mu_{k-1})}\left(\|f\|_\infty\left\|(w_k,\mu_{k-1})-(\hat{w}_k,\tilde{\mu}_k^N)\right\|_p\right.\\
&\quad\left.+\left\|(f\hat{w}_k,\tilde{\mu}_k^N)-(fw_k,\mu_{k-1})\right\|_p\right). \qquad (5.21)
\end{aligned}
$$

Hence, we proceed to finding upper bounds for the two quantities $\|(w_k,\mu_{k-1})-(\hat{w}_k,\tilde{\mu}_k^N)\|_p$ and $\left\|(f\hat{w}_k,\tilde{\mu}_k^N)-(fw_k,\mu_{k-1})\right\|_p$. Note that

$$
\begin{aligned}
&\left\|(f\hat{w}_k,\tilde{\mu}_k^N)-(fw_k,\mu_{k-1})\right\|_p\le\\
&\quad\left\|(fw_k,\mu_{k-1})-(fw_k,\tilde{\mu}_k^N)\right\|_p+\left\|(fw_k,\tilde{\mu}_k^N)-(f\hat{w}_k,\tilde{\mu}_k^N)\right\|_p. \qquad (5.22)
\end{aligned}
$$

For the first term on the right hand side of (5.22), note that $\|w_k\|_p<\infty$, it follows from Lemma 5.4 that

$$
\left\|(fw_k,\mu_{k-1})-(fw_k,\tilde{\mu}_k^N)\right\|_p\le\frac{\tilde{c}_{1,k}\,\|f\|_\infty}{\sqrt{N}}+\|f\|_\infty\,\tilde{d}_{1,k}\sqrt{\Delta}. \qquad (5.23)
$$

For the second term on the right hand side of (5.22), we get using (5.20)

$$
\begin{aligned}
\left|(fw_k,\tilde{\mu}_k^N)-(f\hat{w}_k,\tilde{\mu}_k^N)\right|&=\left|\frac{1}{N}\sum_{i=1}^N f(\tilde{\theta}_k^{(i)})\left(\hat{w}_k^{\tilde{\theta}_k^{(i)}}-w_k^{\tilde{\theta}_k^{(i)}}\right)\right|\\
&\le\frac{\|f\|_\infty}{N}\sum_{i=1}^N\left|\hat{w}_k^{\tilde{\theta}_k^{(i)}}-w_k^{\tilde{\theta}_k^{(i)}}\right|\\
&\le\|f\|_\infty\sup_{1\le i\le N}\left|\hat{w}_k^{\tilde{\theta}_k^{(i)}}-w_k^{\tilde{\theta}_k^{(i)}}\right|\\
&\le\frac{\tilde{c}_{2,k}\,\|f\|_\infty}{\sqrt{N}}+\tilde{d}_{2,k}\,\|f\|_\infty\sqrt{\Delta}.
\end{aligned}
$$

Hence,

$$\left\|(f w_k, \tilde{\mu}_k^N) - (f \hat{w}_k, \tilde{\mu}_k^N)\right\|_p \leq \frac{\tilde{c}_{2,k} \, \|f\|_\infty}{\sqrt{N}} + \tilde{d}_{2,k} \, \|f\|_\infty \sqrt{\Delta}\,.$$

Inserting the latter together with (5.23) in (5.22), we obtain

$$\left\|(f \hat{w}_k, \tilde{\mu}_k^N) - (f w_k, \mu_{k-1})\right\|_p \leq \frac{c_k' \, \|f\|_\infty}{\sqrt{N}} + d_k' \, \|f\|_\infty \sqrt{\Delta}\,, \tag{5.24}$$

where c_k' and d_k' are constants independent of N and Δ. Letting $f = 1$ in (5.24) implies

$$\left\|(w_k, \mu_{k-1}) - (\hat{w}_k, \tilde{\mu}_k^N)\right\|_p \leq \frac{c_k'}{\sqrt{N}} + d_{k'}\sqrt{\Delta}\,. \tag{5.25}$$

Therefore substituting (5.25) and (5.24) into the right hand side of Inequality (5.21), we obtain

$$\left\|(f, \hat{\mu}_k^N) - (f, \mu_k)\right\|_p \leq \frac{\hat{c}_{1,k} \, \|f\|_\infty}{\sqrt{N}} + \hat{d}_{1,k} \, \|f\|_\infty \sqrt{\Delta}\,,$$

where $\hat{c}_{1,k} = \frac{2}{(w_k, \mu_{k-1})} c_k' < \infty$ and $\hat{d}_{1,k} = \frac{2}{(w_k, \mu_{k-1})} d_k' < \infty$ are independent of N and Δ and the statement for $\hat{\mu}_k^N$ follows.

To prove the statement for $\hat{\Gamma}_k^{\tilde{\theta}_k^{(i)}}$, we compute using (2.9) and (4.14)

$$\begin{aligned}
&\left|(f, \hat{\Gamma}_k^{\tilde{\theta}_k^{(i)}}) - (f, \Gamma_k^{\tilde{\theta}_k^{(i)}})\right| \\
&\quad= \left|\frac{((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})}{\hat{w}_k^{\tilde{\theta}_k^{(i)}}} - \frac{((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})}{w_k^{\tilde{\theta}_k^{(i)}}}\right| \\
&\quad= \frac{1}{\hat{w}_k^{\tilde{\theta}_k^{(i)}} \, w_k^{\tilde{\theta}_k^{(i)}}} \left| w_k^{\tilde{\theta}_k^{(i)}} ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}) - \hat{w}_k^{\tilde{\theta}_k^{(i)}} ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right| \\
&\quad\leq \frac{1}{\hat{w}_k^{\tilde{\theta}_k^{(i)}} \, w_k^{\tilde{\theta}_k^{(i)}}} \Bigg(\left| w_k^{\tilde{\theta}_k^{(i)}} ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}) - \hat{w}_k^{\tilde{\theta}_k^{(i)}} ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})\right| \\
&\qquad + \left| \hat{w}_k^{\tilde{\theta}_k^{(i)}} ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}) - \hat{w}_k^{\tilde{\theta}_k^{(i)}} ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}})\right| \Bigg) \\
&\quad\leq \frac{1}{\hat{w}_k^{\tilde{\theta}_k^{(i)}} \, w_k^{\tilde{\theta}_k^{(i)}}} \left(\|f\|_\infty \, \|l_{y_k}\|_\infty \left| w_k^{\tilde{\theta}_k^{(i)}} - \hat{w}_k^{\tilde{\theta}_k^{(i)}} \right| \right) \\
&\qquad + \hat{w}_k^{\tilde{\theta}_k^{(i)}} \left| ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \hat{\pi}_k^{\tilde{\theta}_k^{(i)}}), \hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}}) - ((f l_{y_k}^{\tilde{\theta}_k^{(i)}}, \pi_k^{\tilde{\theta}_k^{(i)}}), \Gamma_{k-1}^{\tilde{\theta}_k^{(i)}}) \right| .
\end{aligned}$$

Since $\hat{w}_k^{\tilde{\theta}_k^{(i)}}, w_k^{\tilde{\theta}_k^{(i)}} \geq \inf_{\theta \in D_\theta} l_{y_k}^{\theta} > 0$, then using Eq. (5.20) and Lemma 5.5 (note that $l_{y_k}^{\tilde{\theta}_k^{(i)}}$ is bounded by Assumption 5.3), we prove the statement of the Lemma. □

5.3 Resampling

In the following lemma we study the convergence of the measure μ_k^N.

Lemma 5.8 *Let the observation sequence $y_{1:k}$ be fixed, for bounded and continuous function f. If*

$$\left\|(f,\hat{\mu}_k^N)-(f,\mu_k)\right\|_p \le \frac{\hat{c}_{1,k}\,\|f\|_\infty}{\sqrt{N}}+\hat{d}_{1,k}\,\|f\|_\infty\sqrt{\Delta}\,,$$

$$\sup_{1\le i\le N}\left|(f,\hat{\Gamma}_k^{\tilde{\theta}_k^{(i)}})-(f,\Gamma_k^{\tilde{\theta}_k^{(i)}})\right| \le \frac{\tilde{c}_{2,k}\,\|f\|_\infty}{\sqrt{N}}+\tilde{d}_{2,k}\,\|f\|_\infty\sqrt{\Delta}\,,$$

holds for some constants $\hat{c}_{1,k}$, $\hat{d}_{1,k}$, $\tilde{c}_{2,k}$ and $\tilde{d}_{2,k}$ which are independent of N and Δ, then there exist constants $c_{1,k}$, $d_{1,k}$, $c_{2,k}$ and $d_{2,k}$ which are independent of N and Δ, such that

$$\left\|(f,\mu_k^N)-(f,\mu_k)\right\|_p \le \frac{c_{1,k}\,\|f\|_\infty}{\sqrt{N}}+d_{1,k}\,\|f\|_\infty\sqrt{\Delta}\,,$$

$$\sup_{1\le i\le N}\left|(f,\hat{\Gamma}_k^{\theta_k^{(i)}})-(f,\Gamma_k^{\theta_k^{(i)}})\right| \le \frac{c_{2,k}\,\|f\|_\infty}{\sqrt{N}}+d_{2,k}\,\|f\|_\infty\sqrt{\Delta}\,.$$

Proof Note that in the resampling step the $\theta_k^{(i)}$ are resampled from the pool $\{\tilde{\theta}_k^{(i)}, i = 1,\cdots,N\}$. Hence it is trivial that

$$\begin{aligned}\sup_{1\le i\le N}\left|(f,\hat{\Gamma}_k^{\theta_k^{(i)}})-(f,\Gamma_k^{\theta_k^{(i)}})\right| &\le \sup_{1\le i\le N}\left|(f,\hat{\Gamma}_k^{\tilde{\theta}_k^{(i)}})-(f,\Gamma_k^{\tilde{\theta}_k^{(i)}})\right|\\ &\le \frac{\tilde{c}_{2,k}\,\|f\|_\infty}{\sqrt{N}}+\tilde{d}_{2,k}\,\|f\|_\infty\sqrt{\Delta}\\ &= \frac{c_{2,k}\,\|f\|_\infty}{\sqrt{N}}+d_{2,k}\,\|f\|_\infty\sqrt{\Delta}\,,\end{aligned}$$

where $c_{2,k} = \tilde{c}_{2,k}$ and $d_{2,k} = \tilde{d}_{2,k}$ are independent of N and Δ. Moreover, by triangle inequality, we have

$$\left\|(f,\mu_k^N)-(f,\mu_k)\right\|_p \le \left\|(f,\mu_k^N)-(f,\hat{\mu}_k^N)\right\|_p+\left\|(f,\hat{\mu}_k^N)-(f,\mu_k)\right\|_p\,. \tag{5.26}$$

For the second term on the right hand side of (5.26), it follows from the conditions in the lemma, that

$$\left\|(f,\hat{\mu}_k^N)-(f,\mu_k)\right\|_p \le \frac{\hat{c}_{1,k}\,\|f\|_\infty}{\sqrt{N}}+\hat{d}_{1,k}\,\|f\|_\infty\sqrt{\Delta}\,. \tag{5.27}$$

Note that $\{\theta_k^{(i)}, i = 1,\cdots,N\}$ are i.i.d. samples generated from $\hat{\mu}_k^N(\mathrm{d}\theta) = \sum_{i=1}^N \hat{w}_k^{\tilde{\theta}_k^{(i)}}\,\delta_{\tilde{\theta}_k^{(i)}}(\mathrm{d}\theta)$. Let $\tilde{\mathcal{G}}_k$ be the sigma-algebra generated by $\{\theta_{1:k-1}^{(i)}, \tilde{\theta}_{1:k}^{(i)}, i = 1,\cdots,N\}$, then

$$\begin{aligned}\mathbb{E}[f(\theta_k^{(i)})\mid\tilde{\mathcal{G}}_k] &= \int f(\theta)\sum_{i=1}^N \hat{w}_k^{\tilde{\theta}_k^{(i)}}\,\delta_{\tilde{\theta}_k^{(i)}}(d\theta)\\ &= \sum_{i=1}^N \hat{w}_k^{\tilde{\theta}_k^{(i)}}\,f(\tilde{\theta}_k^{(i)})\\ &= (f,\hat{\mu}_k^N)\,.\end{aligned}$$

Define $Z_k^{(i)} = f(\theta_k^{(i)}) - (f,\hat{\mu}_k^N) = f(\theta_k^{(i)}) - \mathbb{E}[f(\theta_k^{(i)})\mid\tilde{\mathcal{G}}_k]$. The $Z_k^{(i)}, i = 1,\cdots,N$ are random variables with zero-mean and bounded by $2\,\|f\|_\infty$ and have the property $\mathbb{E}[Z_k^{(i)}Z_k^{(j)}\mid$

$\tilde{\mathcal{G}}_k] = 0$ for $i \neq j$. Let $p \geq 1$, and let's first additionally assume p is an even integer. Then

$$\begin{aligned}\mathbb{E}\left[\left|(f, \mu_k^N) - (f, \hat{\mu}_k^N)\right|^p \mid \tilde{\mathcal{G}}_k\right] &= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(\theta_k^{(i)}) - (f, \hat{\mu}_k^N)\right|^p \mid \tilde{\mathcal{G}}_k\right] \\ &= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N Z_k^{(i)}\right|^p \mid \tilde{\mathcal{G}}_k\right] \\ &= \frac{1}{N^p}\mathbb{E}\left[\sum_{i_1=1}^N \cdots \sum_{i_p=1}^N Z_k^{(i_1)} \cdots Z_k^{(i_p)} \mid \tilde{\mathcal{G}}_k\right].\end{aligned}$$

Since $\mathbb{E}[Z_k^{(i)} \mid \tilde{\mathcal{G}}_k] = 0$, there are at most $N^{\frac{p}{2}}$ non-zero contributions to

$$\sum_{i_1=1}^N \cdots \sum_{i_p=1}^N \mathbb{E}\left[Z_k^{(i_1)} \cdots Z_k^{(i_p)} \mid \tilde{\mathcal{G}}_k\right].$$

Hence

$$\mathbb{E}\left[\sum_{i_1=1}^N \cdots \sum_{i_p=1}^N Z_k^{(i_1)} \cdots Z_k^{(i_p)} \mid \tilde{\mathcal{G}}_k\right] \leq \frac{1}{N^{\frac{p}{2}}} 2^p \left\|f\right\|_\infty^p ,$$

which implies

$$\left\|(f, \mu_k^N) - (f, \hat{\mu}_k^N)\right\|_p \leq \frac{2\left\|f\right\|_\infty}{\sqrt{N}}. \tag{5.28}$$

Substituting Inequality (5.28) and (5.27) into Eq. (5.26) yields, for any even p, the result

$$\left\|(f, \mu_k^N) - (f, \mu_k)\right\|_p \leq \frac{c_{1,k}\left\|f\right\|_\infty}{\sqrt{N}} + d_{1,k}\left\|f\right\|_\infty \sqrt{\Delta}, \tag{5.29}$$

where $c_{1,k} = 2\left\|f\right\|_\infty + \hat{c}_{1,k}$ and $d_{1,k} = \hat{d}_{1,k}$. For any real number $p \geq 1$, we know there exist an even number $q > p$ such that (5.29) holds for this number q. Hence the statement is proved. □

5.4 Convergence of the Kalman Particle algorithm

In the following theorem we prove the convergence of our proposed algorithm.

Theorem 5.9 *Suppose the function f is bounded and continuous and Assumptions 4.1, 5.1, 5.2, 5.3 hold. Let the sequence of the observation $y_{1:k}$ be fixed and the measures μ_k and μ_k^N resulting from Algorithm 4.2 be respectively as in (5.1) and (5.4), where the model is Gaussian and linear in the sense of (2.11) or is of type (3.1). Then it holds*

$$\left\|(f, \mu_k^N) - (f, \mu_k)\right\|_p \leq \frac{c_{1,k}\left\|f\right\|_\infty}{\sqrt{N}} + d_{1,k}\left\|f\right\|_\infty \sqrt{\Delta},$$

with constants $c_{1,k}, d_{1,k}$, $1 \leq k \leq K$ independent of N and Δ.

Proof We prove this theorem by induction. At time t_0, the parameter samples $\theta_0^{(i)}$, $i = 1, \cdots, N$, are sampled from the initial measure μ_0 and $\mu_0^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_0^{(i)}}$. A well known result of Monte Carlo simulation, see for example Chapter I in Glasserman (2013), implies that

$$\left\|(f,\mu_0)-(f,\mu_0^N)\right\|_p \le \frac{c_{1,t_0}\,\|f\|_\infty}{\sqrt{N}} + d_{1,t_0}\,\|f\|_\infty\sqrt{\Delta},$$

for some constant c_{1,t_0} and d_{1,t_0} independent of N and Δ. Moreover, one can just define the initial measure of x_0 by a Gaussian measure Γ_0 and define $\hat{\Gamma}_0^{\theta_0^{(i)}} = \Gamma_0^{\theta_0^{(i)}} = \Gamma_0$, for $i = 1, \cdots, N$. Then it is trivial that

$$\sup_{1\le i\le N}\left|(f,\hat{\Gamma}_0^{\theta_0^{(i)}})-(f,\Gamma_0^{\theta_0^{(i)}})\right| \le \frac{c_{2,t_0}\,\|f\|_\infty}{\sqrt{N}} + d_{2,t_0}\,\|f\|_\infty\sqrt{\Delta}$$

holds for some constant c_{2,t_0} and d_{2,t_0} independent of N and Δ (actually $c_{2,t_0} = d_{2,t_0} = 0$). Assume that, at time $k-1$, the inequalities

$$\left\|(f,\mu_{k-1}^N)-(f,\mu_{k-1})\right\|_p \le \frac{c_{1,k-1}\,\|f\|_\infty}{\sqrt{N}} + d_{1,k-1}\,\|f\|_\infty\sqrt{\Delta}$$

and

$$\sup_{1\le i\le N}\left|(f,\hat{\Gamma}_{k-1}^{\theta_{k-1}^{(i)}})-(f,\Gamma_{k-1}^{\theta_{k-1}^{(i)}})\right| \le \frac{c_{2,k-1}\,\|f\|_\infty}{\sqrt{N}} + d_{2,k-1}\,\|f\|_\infty\sqrt{\Delta}$$

hold for some constants $c_{1,k-1}$, $d_{1,k-1}$, $c_{2,k-1}$ and $d_{2,k-1}$ which are independent of N and Δ. Then we can just successively apply Lemmas 5.7 and 5.8 to obtain the statements of the lemma. □

6 Numerical results on affine term structure models

In this section we illustrate our results by considering some specific examples of the affine class (3.1) to which we apply our algorithms. In particular we consider the one-factor Cox-Ingersoll-Ross (CIR) model, the two-factor Vasiček model, and the one-factor Vasiček model with stochastic volatility. To test how the designed algorithm works on these models, the models are calibrated on simulated data using pre-determined parameters. Moreover, for the illustration of our algorithm on empirical data, we consider the calibration of a two-factor Vasiček model on yield curves. We compare the behavior of our algorithm to the one generated by the recursive nested particle filter (RNPF) for the CIR model. The comparison shows that in this case our algorithm outperforms the RNPF.

6.1 One-factor CIR model

At first we consider the CIR model (3.6). One important property of the CIR models is that the yield curves are always non-negative. The transition density of the CIR model has a non-central chi-square distribution, i.e.,

$$x_{t_{k+1}} \mid x_{t_k} \sim c\chi_p^2(\lambda)\,,$$

where $p = \frac{4\alpha\beta}{\sigma^2}$ is the degree of freedom, $\lambda = x_k(4\alpha e^{-\alpha(t_{k+1}-t_k)})/(\sigma^2(1-e^{-\alpha(t_{k+1}-t_k)}))$ is the non-centrality parameter and $c = \sigma^2(1-e^{-\alpha(t_{k+1}-t_k)})/4\alpha$. We define the short rate process to be $r_t = x_t$. The analytical solution of the functions ϕ and ψ defined in (3.2) can be obtained by solving the ODE (A.2), with $d = 1$, $\gamma = -1$ and $c = 0$.

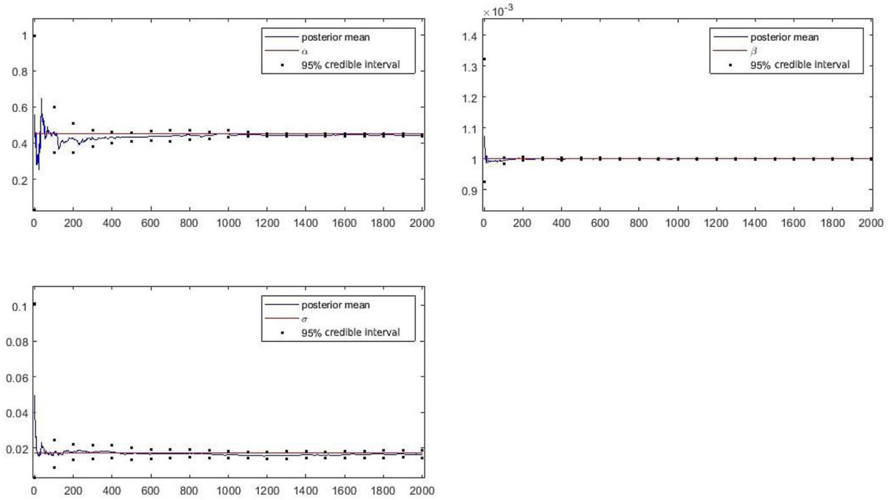

Fig. 1 Convergence of parameter estimates by KPF for the CIR model with noise variance $h = 1 \times 10^{-8}$; the blue lines represent the posterior mean of the parameters, the red lines represent the true values of the parameters and the black dots are the lower and upper bounds of the 95% credible intervals

In the tests, the parameters are set to be $\alpha_1 = 0.45$, $\beta = 0.001$ and $\tilde{\Sigma} = 0.017$. Based on these parameters values, we generate daily data for yield curves with the times to maturity ranging from 1 year to 30 years. The time length is $T = 2000$, i.e., the data set contains 2000 days. Furthermore, we add white noise with variance $h = 1 \times 10^{-8}$ in the simulated zero rates.

We use our Kalman particle filter algorithm (Algorithm 4.2, henceforth referred to as KPF) to calibrate the model parameters. The transition distribution of the CIR model is not Gaussian, but note that the CIR model fits the structure of (3.1) with $\Sigma = 0$. Hence we use the approximation (4.10) and apply the Kalman filter in the inner layer. In the outer layer, we set the number of particles to be $N = 5000$ and the initial prior distribution of the parameters to be uniform, i.e.,

$$\alpha \sim U(0, 1), \quad \beta \sim U(0, 0.01), \quad \text{and} \quad \sigma \sim U(0, 0.1)\,.$$

Moreover, the sampled parameters at each step are truncated by the boundary of the latter corresponding uniform distributions respectively. In the next sections we use the same kind of truncation. In the inner layer, the mean and variance of the initial prior distribution of x_0 are 0.005 and 0.01. The discounting factor a is set to be 0.98 and the variance boundaries V_N and V_f are set to be the diagonal matrices whose elements are equal to $\frac{1}{\sqrt{N^3}}$, i.e. $p = 1$ and 10^{-8}, respectively. We will also use the same a, V_N and V_f for other experiments later.

Figure 1 shows that the estimated parameters converge over time to the real value. In this figure (and in other ones below) the blue lines represent the posterior mean of the parameters, the red lines represent the true values of the parameters and the black dots are the lower and upper bounds of the 95% credible intervals.

Figure 2 shows the estimated states follow the simulated states very well. Moreover, we observe that after 810 steps the variance is smaller than the required level V_N. We repeated this experiment for noise variance $h = 1 \times 10^{-7}$ and $h = 1 \times 10^{-9}$ and the results are shown in Figures 3 and 4 , respectively.

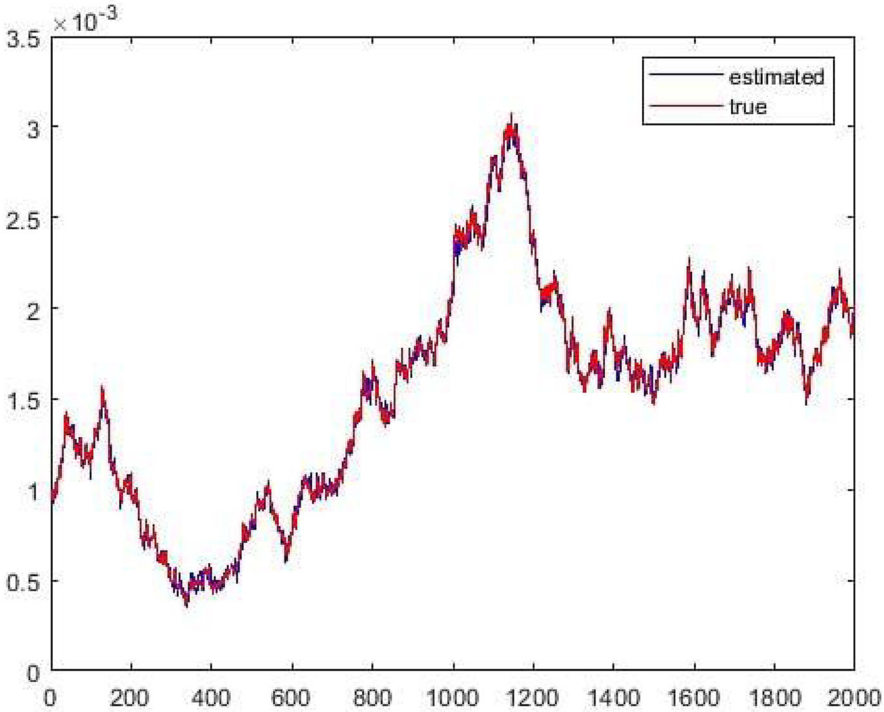

Fig. 2 Simulated states and estimated states; the blue line represents estimates of the simulated states, the red line represents the simulated states

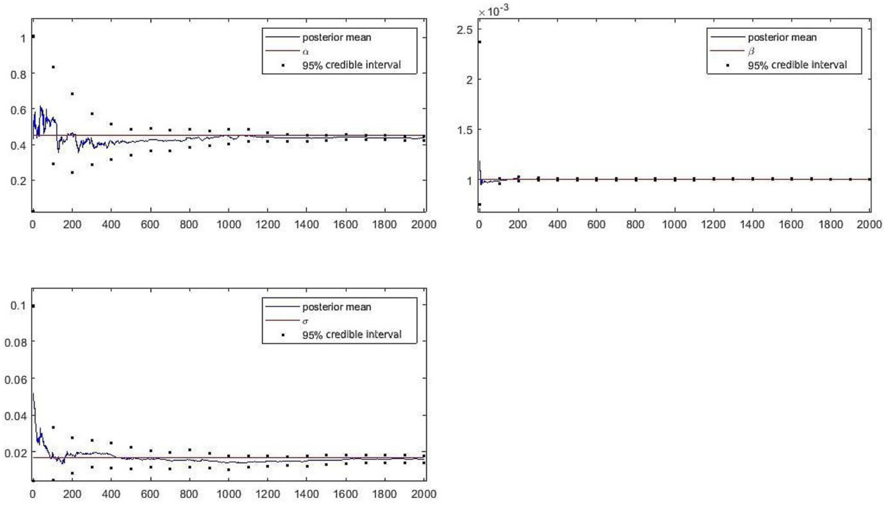

Fig. 3 Convergence of parameter estimates by KPF for the CIR model with noise level $h = 1 \times 10^{-7}$

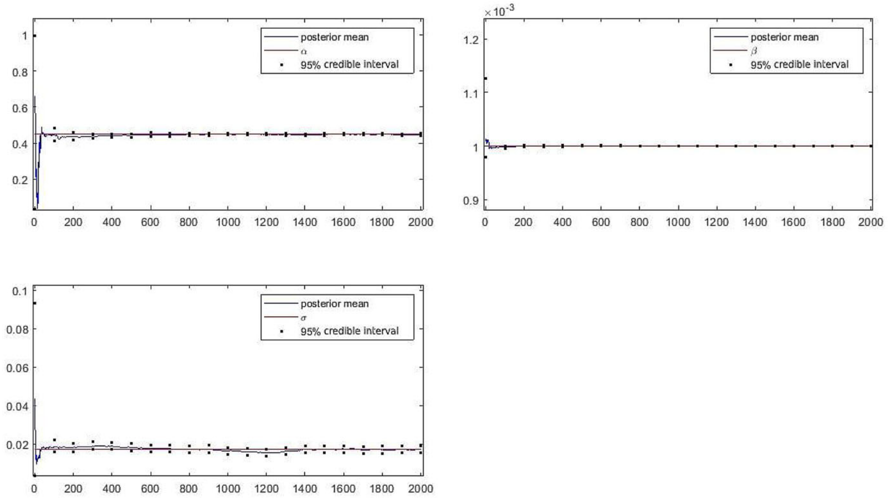

Fig. 4 Convergence of parameter estimates by KPF for the CIR model with noise level $h = 1 \times 10^{-9}$

We find out that in all these cases the posterior means converge to the real value. We also observed that the bigger the variance of the noise is, the more steps it takes to switch the jittering kernel and the more steps it takes to the algorithm to converge. Specifically, in these three cases with noise levels 1×10^{-7}, 1×10^{-8} and 1×10^{-9}, it has taken 1164, 810 and 134 steps to switch between the kernels, respectively.

Furthermore, for noise variance $h = 1 \times 10^{-8}$, we extend this experiment to weekly and monthly data, i.e. the time step is weekly and monthly and the results are shown in Figures 5 and 6 , respectively. We find out that also in both cases the posterior means converge to the real value. We also observe that the bigger the time step is, the more time it takes for the algorithm to converge (note that 1 step in those two cases is 1 week and 1 month respectively).

We also implemented the recursive nested particle filter (RNPF) for comparison. The sample size in the two layers are set to be 1000 and 300 respectively. Moreover, we simulated 18000 more days of data, hence in total we obtain 20000 days of data. For the rest we use the same settings as in the previous example, i.e. the same initial sample distributions for the parameter generation, the same boundary on the parameter samples in the outer layer, and the variance of the jittering kernel is also set the same as V_N.

Figure 7 shows how the behavior of the estimated parameters over time. One observes that even after 20000 time steps the RNPF algorithm estimates of α and σ do not reach the correct parameter values.

6.2 Two-factor Vasicek model

In this subsection we consider the two-factor Vasiček model (3.7). The short rate process r_t is given by $r_t = x_t^{(1)} + x_t^{(2)}$. The analytical solution of the functions ϕ and ψ defined in (3.2) can be obtained by solving the ODE (A.1), with $d = 2$, $\gamma = (-1, -1)^\top$ and $c = 0$. Furthermore we set $\alpha_{11} = 0.03$, $\alpha_{22} = 0.23$, $\Sigma_{11} = 0.02$, $\Sigma_{12} = 0$, $\Sigma_{21} = 0.02\rho$, $\Sigma_{22} = 0.02\sqrt{1-\rho^2}$, $\rho = -0.5$. Similar as in the previous example, the yield curve data are simulated based on these parameters and then a white noise process is added on the simulated data. The variance of the white noise is 6×10^{-7}. The times to maturity of the

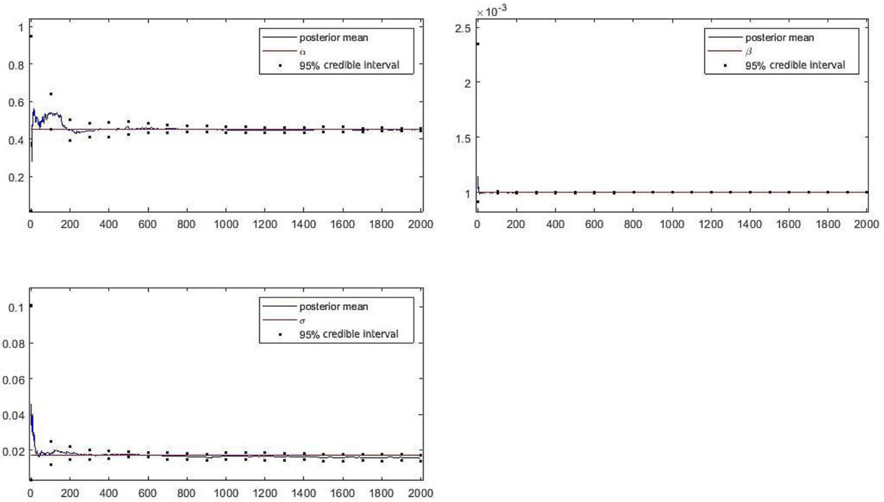

Fig. 5 Convergence of parameter estimates by KPF for the CIR model with weekly data and noise level $h = 1 \times 10^{-8}$

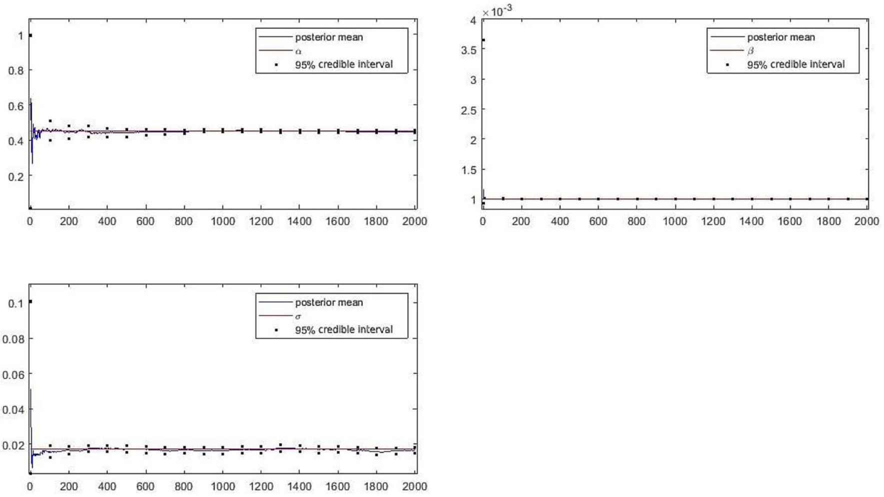

Fig. 6 Convergence of parameter estimates by KPF for the CIR model with monthly data and noise level $h = 1 \times 10^{-8}$

yield curves range from 1 year up to 30 years. The time step of the data is set to be daily and the time length is 2000.

Using these noisy simulated data, we use our Kalman particle algorithm 4.2 to calibrate the model parameters. The Vasiček fits the structure of (3.1) with $\Sigma = 0$. Since the Vasiček model is a Gaussian model, the Kalman filter in the inner layer gives an optimal filter. The number of particles at each step is $N = 2000$. For the initialization of the parameter samples, the initial prior distribution of parameters is chosen to be uniform, namely:

$$\alpha_1, \alpha_2 \sim U(0, 0.4), \quad \sigma_1, \sigma_2 \sim U(0, 0.1), \quad \text{and} \quad \rho \sim U(-0.8, -0.3) .$$

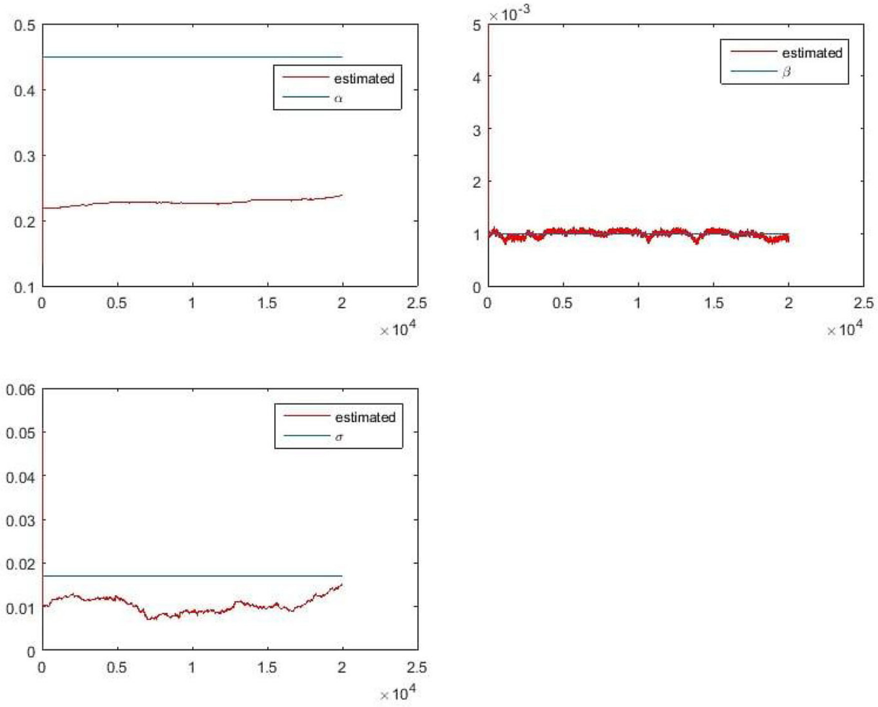

Fig. 7 Behavior of parameter estimates by RNPF for the CIR model

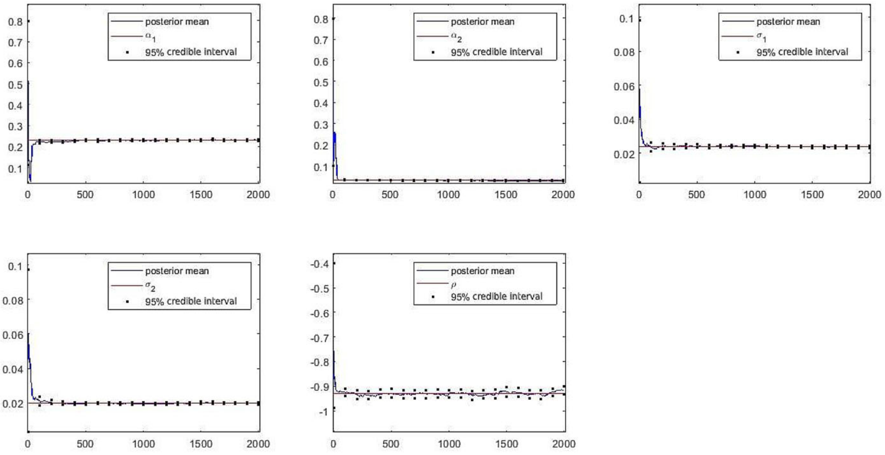

Fig. 8 Convergence of parameter estimates by KPF for the Vasiček model

The initial prior distribution of the state x_k is chosen to be Gaussian with mean 0 and variance diag$([0.1, 0.1])$. Figure 8 shows the how the estimated parameters converge over time. One can observe that the convergence is very fast and accurate.

6.3 Vasicek model with stochastic volatility

We consider the following stochastic volatility model,

$$
\begin{aligned}
\mathrm{d}V_t &= \alpha_1(0.1 - V_t)\,\mathrm{d}t + \sigma_1\sqrt{V_t}\,\mathrm{d}W_t^{(1)},\\
\mathrm{d}x_t &= \alpha_2(\beta - x_t)\,\mathrm{d}t + \sigma_2\sqrt{V_t}\left(\rho\,\mathrm{d}W_t^{(1)} + \sqrt{1-\rho^2}\,\mathrm{d}W_t^{(2)}\right),
\end{aligned}
$$

which is model (3.8) in a notation that is more suitable for the purposes of this section. The process V_t presents the fluctuation of the volatility of the system. Note that the long term mean of the process V_t is fixed at the known constant 0.1, otherwise the model would be over-parametrized, i.e. by scaling the volatility process and the parameters σ_1, σ_2 one can obtain an equivalent model. The transition density of this stochastic volatility model is not analytically available. To tackle this issue, we apply the same approximation as in the CIR model test of Sect. 6.1, namely, we approximate the stochastic diffusion by constant diffusion between the time steps, see (4.10). Under this approximation, the transition density of the model is Gaussian and the mean and variance can be theoretically computed. The short rate is defined by $r_t = x_t$ and hence the yield curve can be computed, see (3.3), from $R_t(\tau) = \phi(\tau, 0) + \psi(\tau, 0)^\top \tilde{x}_t$, with $\tilde{x}_t = [V_t, x_t]^\top$. The functions ϕ, ψ are the solutions to the Riccati Eqs. (A.2), with $d = 2$, $\gamma = (0, -1)^\top$ and $c = 0$. The solutions to these latter equations are not known in closed form. We introduce an efficient numerical algorithm to compute these functions, the detailed algorithm is in the Appendices A.3. The parameters of this model are set to be $[\alpha_1, \alpha_2, \beta, \sigma_1, \sigma_2, \rho] = [0.1, 0.3, 0.03, 0.3, 0.07, -0.5]$. The variance of the white noise is 10^{-8}. The times to maturity of the yield curves range from 1 year up to 20 years. The time step of the data is set to be daily and the time length is 2000. In the outer layer, we sample $N = 2000$ particles and in the inner layer, we use the Kalman filter. The initial prior distributions of the parameters are uniform,

$$
\begin{aligned}
&\alpha_1, \alpha_2 \sim U(0, 1), \beta \sim U(0, 0.1),\\
&\sigma_1 \sim U(0, 0.8), \sigma_2 \sim U(0, 0.2) \text{ and } \rho \sim U(-1, 1).
\end{aligned}
$$

The mean and variance of the initial prior distribution of $\tilde{x}_t$ are $[0.1, 0]$ and $\mathrm{diag}([0.01, 0.01])$. Figure 9 shows that also for a model with stochastic volatility the parameter estimates quickly converge.

6.4 One-factor CIR model with jump parameters

In this section we test the performance of Algorithm 4.3 in case a sudden jump in the parameter takes place, as the alternative of Algorithm 4.2 for the case of static parameters. This experiment can be seen as an extension of the experiment on the CIR model calibration of Sect. 6.1. In this experiment, we assume the parameters have a jump at time $T = 2001$, from $[\alpha, \beta, \sigma] = [0.45, 0.001, 0.017]$ to $[\alpha, \beta, \sigma] = [0.55, 0.0015, 0.023]$. We simulate the new data from time point $T = 2001$ to $T = 4000$ based on the new parameters and the settings for the other parameters are the same as in Sect. 6.1. To identify the parameter change, we set $q = 0.1$. Figure 10 shows that in this study the KPF algorithm for models with time-varying parameters is able to track a sudden change in the parameter values and quickly stabilizes at the new values.

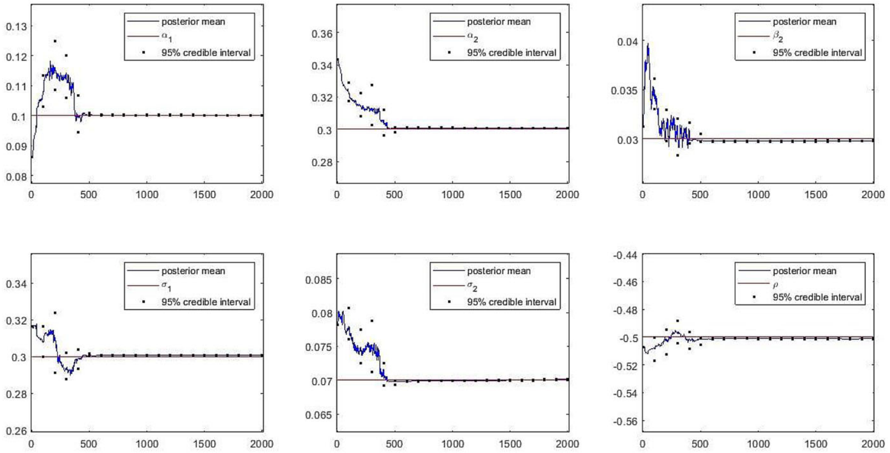

Fig. 9 Convergence of parameter estimates by KPF for the Vasiček model with stochastic volatility

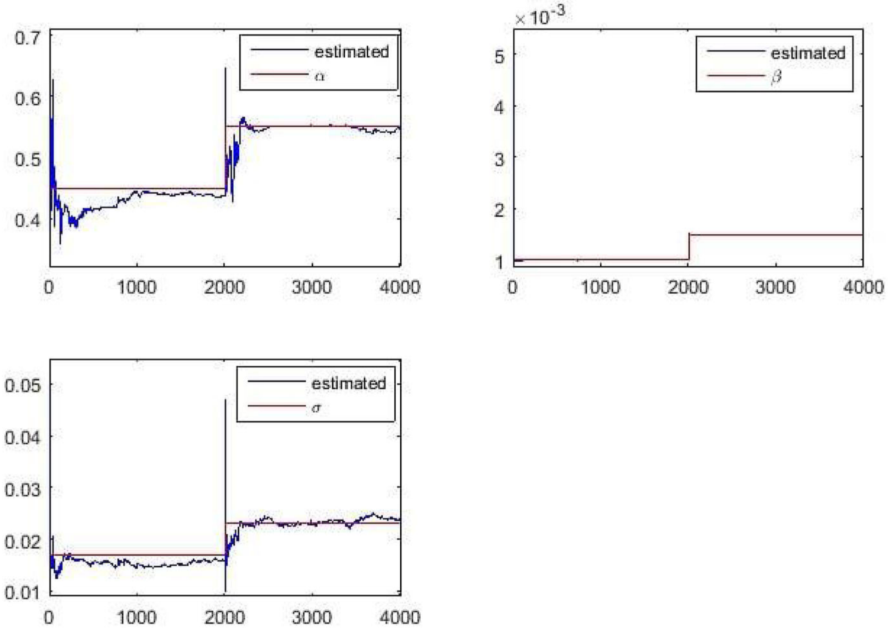

Fig. 10 Parameter estimates by KPF for the CIR model with a sudden jump

6.5 Two-factor Vasicek model for real data

In this subsection, we apply our algorithm to the calibration of a two-factor Vasiček model on yield curves that are bootstrapped from Euro swap rates with 3-month floating and fixed leg. The swap rates come from Bloomberg, with daily time step from 01-01-2009 to 29-05-2017. The tenor we use are from 4 years to 15 years.

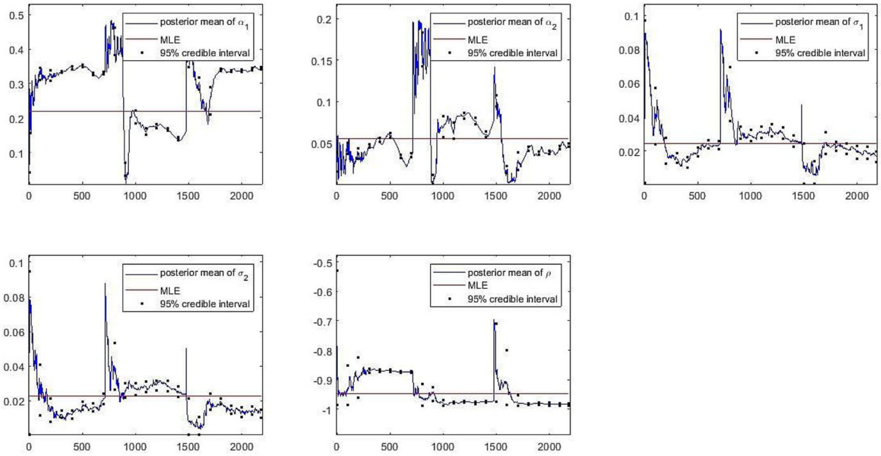

Fig. 11 Parameter estimates by KPF and MLE for the two-factor Vasiček model applied to real data of yield curves

Let $(Y_t)_{t\geq 0} = (Y_t^{(1)}, Y_t^{(2)})_{t\geq 0}$ be a two-factor affine process. Assume the short rate $r_t = Y_t^{(1)} + Y_t^{(2)}$, $t \geq 0$. Then the corresponding yield curves with tenor τ is given by

$$R_t(\tau) = \phi(\tau, 0) + \psi(\tau, 0)^\top Y_t.$$

We suppose we have observations of the process Y at discrete times $k = 1, \dots T$. Denote by $\bar{R}(\tau) = \frac{1}{T}\sum_{k=1}^{T} R_k(\tau)$ and $\bar{Y} = \frac{1}{T}\sum_{k=1}^{T} Y_k$. Then we obtain

$$R_k(\tau) - \bar{R}(\tau) = \psi(\tau, 0)^\top (Y_k - \bar{Y})\,. \tag{6.1}$$

We assume the process $X_t = Y_t - \bar{Y}$, $t \geq 0$, is a two-factor Vasiček process with mean reversion level zero, i.e.,

$$\mathrm{d}X_t = AX_t\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t\,, \tag{6.2}$$

where A is a 2×2 diagonal matrix and Σ is a 2×2 lower diagonal matrix. Instead of the perfect observations given by (6.1), we assume to have noisy observations of the yield curve given by

$$R_k(\tau) - \bar{R}(\tau) = \psi(\tau, 0)^\top X_k + v_k, \tag{6.3}$$

which corresponds to Eq. (3.5). Moreover, $\psi(\tau, 0)$ the solution of ODE (A.1), with $d = 2$, $\gamma = (-1, -1)^\top$ and $c = 0$.

The model (6.2) has five parameters that need to be calibrated: the diagonal elements of A, denoted by α_1, α_2, the volatility parameters and the correlation parameter in Σ, denoted by σ_1, σ_2 and ρ, respectively. The variance of v_k in (6.3) is set to be 2.36×10^{-8} (from MLE). The base configuration of our algorithm is the same as in Sect. 6.2, except that we now also allow for sudden changes of the parameters. For that we set $q = 0.01$. Figure 11 shows for all five parameters the calibration results using the KPF algorithm and the MLE. Figure 11 shows that there are three different levels of the parameters to which the algorithm quickly converges in all cases.

7 Conclusion

In this paper we have introduced a semi-recursive algorithm combining the Kalman filter and the particle filter with a two layers structure. In the outer layer the dynamic Gaussian kernel is implemented to sample the parameter particles. Moreover, the Kalman filter is applied the inner layer to estimate the posterior distribution of the state variables given the parameters sampled in the outer layer. These two changes provide faster convergence and reduce the computational time comparable to the RNPF methodology. The theoretical contribution of this paper is the convergence analysis of the proposed algorithm. We proved that, under regularity assumptions and given a certain model structure, the posterior distribution of the parameters and the state variables converge to the actual distribution in L^p with rate $\mathcal{O}(N^{-\frac{1}{2}} + \Delta^{\frac{1}{2}})$. The theoretical result is complemented by numerical results for several affine term structure models with static parameters or jump parameters. Although our numerical illustrations are for term structure models, the Kalman particle algorithm can also be applied to many other models.

Acknowledgements The authors thank the contribution from Dr. Peter den Iseger and the support from ABN AMRO Bank N.V. Two anonymous referees are gratefully acknowledged for their constructive suggestions for improvements.

A Affine processes

Affine processes are continuous-time Markov processes characterised by the fact that their characteristic function depends in an exponentially affine way on the initial state vector of the process. From Theorem 2.7 in Duffie et al. (2003), we know that the model of type (3.1) is an $\mathbb{R}_+^p \times \mathbb{R}^q$-valued affine process given Σ or $\tilde{\Sigma}$ is zero and the admissibility of the parameters of this model.

A.1 Admissibility of the parameters

Let $\Gamma = \Sigma\Sigma^\top$, $\tilde{\Gamma} = \tilde{\Sigma}\tilde{\Sigma}^\top$, $I = \{1, \cdots, p\}$ and $J = \{p+1, \cdots, p+q\}$. Here below, we introduce the admissibility of the parameters of (3.1) when $\tilde{\Sigma} = 0$,

- $\Gamma_{II} = 0$,
- $A\beta \in \mathbb{R}_+^p \times \mathbb{R}^q$,
- $A_{IJ} = 0$,
- A_{II} has nonpositive off-diagonal elements .

When $\Sigma = 0$, then the admissibility of the parameters of (3.1) reads

- $\Gamma = 0$, if $p = 0$,
- $\Gamma_{kl} = \Gamma_{lk} = 0$, for $k \in I/\{1\}$, for all $1 \leq l \leq d$,

- $A\beta \in \mathbb{R}_+^p \times \mathbb{R}^q$,
- $A_{IJ} = 0$,
- A_{II} has nonpositive off-diagonal elements .

This latter conditions on the parameters ensure that the process of (3.1) remains in the state space $\mathbb{R}_+^p \times \mathbb{R}^q$.

A.2 Riccati equations

Let $u \in \mathbb{C}^d$, $(\phi(\cdot, u), \psi(\cdot, u)) : [0, T] \to \mathbb{C} \times \mathbb{C}^d$ and $(\tilde{\phi}(\cdot, u), \tilde{\psi}(\cdot, u)) : [0, T] \to \mathbb{C} \times \mathbb{C}^d$ be C^1-functions. We introduce the following generalised Riccati equations.

Generalised Riccati equations (1).

$$
\begin{aligned}
\partial_t \phi(t, u) &= \frac{1}{2} \psi_J^\top(t, u) \Gamma_{JJ} \psi_J(t, u)) + (A\beta)^\top \psi(t, u) - c\,, \\
\phi(0, u) &= 0\,, \\
\partial_t \psi_i(t, u) &= -A_i^\top \psi(t, u) - \gamma_i\,, \quad 1 \le i \le d\,, \\
\partial_t \psi_J(t, u) &= -A_{JJ} \psi(t, u) - \gamma_J\,, \\
\psi(0, u) &= u\,,
\end{aligned}
\tag{A.1}
$$

Generalised Riccati equations (2).

$$
\begin{aligned}
\partial_t \tilde{\phi}(t, u) &= (A\beta)^\top \tilde{\psi}(t, u) - c\,, \\
\tilde{\phi}(0, u) &= 0\,, \\
\partial_t \tilde{\psi}_1(t, u) &= \frac{1}{2} \tilde{\psi}^\top(t, u) \tilde{\Gamma}_1^\top \tilde{\psi}(t, u) - A_1^\top \tilde{\psi}(t, u) - \gamma_1\,, \\
\partial_t \tilde{\psi}_i(t, u) &= -A_i^\top \tilde{\psi}(t, u) - \gamma_i\,, \quad 2 \le i \le d\,, \\
\partial_t \tilde{\psi}_J(t, u) &= -A_{JJ}^\top \tilde{\psi}_J(t, u) - \gamma_J\,, \\
\tilde{\psi}(0, u) &= u\,.
\end{aligned}
\tag{A.2}
$$

The aim in the following theorem is to compute the zero coupon bond price $P(t, T)$ introduced in (3.2). For a proof, we refer to Theorem 3.1 in Keller-Ressel and Mayerhofer (2015).

Theorem A.1 *Let $\tau > 0$ and $(x_t)_{t \ge 0}$ be as in (3.1) with $\tilde{\Sigma} = 0$. Then the following statements are equivalent*

1. $\mathbb{E}[\mathrm{e}^{-\int_0^\tau r(s)\,\mathrm{d}s}] < \infty$*, for some* $x \in \mathbb{R}_+^p \times \mathbb{R}^q$.
2. *There exists a unique solution* (ϕ, ψ) *on* $[0, \tau]$ *to the generalised Riccati equations (A.1) with initial data* $u = 0$.

In any of the above cases, it holds for all $0 \le t \le T \le \tau$ *and for all* $x \in \mathbb{R}_+^p \times \mathbb{R}^q$,

$$
\mathbb{E}[\mathrm{e}^{-\int_t^T r(s)\,\mathrm{d}s} \mid \mathcal{F}_t] = \mathrm{e}^{-\phi(T-t,0) - \psi(T-t,0)x(t)}\,.
$$

The above statements remain true when $\Sigma = 0$ *if we replace* (ϕ, ψ) *by* $(\tilde{\phi}, \tilde{\psi})$*, the solution to (A.2).*

A.3 Numerical solution to Riccati equations

For many Riccati equations, it is hard (or even impossible) to calculate a closed-form solution, especially in high dimensional cases. So a numerical approach is needed. In general, the Riccati equations for $\phi(t,u)$ and $\psi(t,u)$ are given by

$$\begin{aligned}
\partial_t\phi(t,u) &= \frac{1}{2}\psi(t,u)^\top a\psi(t,u) + b^\top\psi(t,u) - c,\\
\phi(0,u) &= 0;\\
\partial_t\psi_i(t,u) &= \frac{1}{2}\psi(t,u)^\top\alpha_i\psi(t,u) + \beta_i^\top\psi(t,u) - \gamma_i,\\
\psi(0,u) &= u.
\end{aligned} \tag{A.3}$$

with known parameters $a, b, c, \alpha_i, \beta_i$ and γ_i, $i = 1, \cdots, d$.

We use a Taylor series to approximate the solution (ϕ, ψ). In order to do so, first we need to determine the coefficients in Taylor expansion.

Proposition A.2 *Suppose $(\phi(t,u), \psi(t,u))$ is the solution of (A.3). Given the value of u, assume the Taylor expansions of $(\phi(t,u), \psi(t,u))$ are given by $\phi(t,u) = \sum_{k=0}^{\infty} C_k(u)t^k < \infty$, and $\psi_i(t,u) = \sum_{k=0}^{\infty} D_k^i(u)t^k < \infty$. Then we have the following recursion for the coefficients:*

$$\begin{aligned}
C_0(u) &= 0,\\
C_1(u) &= \frac{1}{2}u^\top au + b^\top u - c,\\
C_{k+1}(u) &= \frac{1}{1+k}\left(\frac{1}{2}\sum_{n=0}^{k} D_n^\top(u)aD_{k-n}(u) + b^\top D_k u\right), k \geq 2,\\
D_0^i(u) &= u_i,\\
D_1^i(u) &= \frac{1}{2}u^\top\alpha_i u + \beta_i^\top u - \gamma_i,\\
D_{k+1}^i(u) &= \frac{1}{1+k}\left(\frac{1}{2}\sum_{n=0}^{k} D_n^\top(u)\alpha_i D_{k-n}(u) + \beta_i^\top D_k(u)\right), k \geq 2,
\end{aligned}$$

where $D_k(u) = (D_k^1(u), \cdots, D_k^d(u))^\top$, $k = 0, \cdots, n$.

Proof Suppose $\phi(t,u) = \sum_{k=0}^{\infty} C_k(u)t^k$, $\psi_i(t,u) = \sum_{k=0}^{\infty} D_k^i(u)t^k$, let $t = 0$, we obtain $C_0(u) = 0$, $D_0^i(u) = u_i$. Taking the derivative of $\psi_i(t,u)$ w.r.t. t,

$$\begin{aligned}
\partial_t\psi_i(t,u) &= \sum_{k=1}^{\infty} D_k^i(u)kt^{k-1}\\
&= \sum_{k=0}^{\infty} D_{k+1}^i(u)(k+1)t^k.
\end{aligned} \tag{A.4}$$

On the other hand, according to (A.3),

$$\partial_t\psi_i(t,u) = \frac{1}{2}\psi(t,u)^\top\alpha_i\psi(t,u) + \beta_i^\top\psi(t,u) - \gamma_i$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{l,r=1}^{d} \psi_l(t,u)\alpha_i(l,r)\psi_r(t,u) + \sum_{s=1}^{d} B(s,i)\psi_s(t,u) - \gamma_i \\
&= \frac{1}{2}\sum_{l,r=1}^{d} (\sum_{k=0}^{\infty} D_k^l(u)t^k)\alpha_i(l,r)(\sum_{k=0}^{\infty} D_k^r(u)t^k) + \sum_{s=1}^{d} B(s,i)(\sum_{k=0}^{\infty} D_k^s(u)t^k) - \gamma_i \\
&= \frac{1}{2}\sum_{l,r=1}^{d} \alpha_i(l,r)\sum_{k=0}^{\infty}(\sum_{m=0}^{k} D_m^l(u)D_{k-m}^r(u))t^k + \sum_{s=1}^{d} B(s,i)(\sum_{k=0}^{\infty} D_k^s(u)t^k) - \gamma_i \\
&= \sum_{k=0}^{\infty}\left(\frac{1}{2}\sum_{l,r=1}^{d}\sum_{m=0}^{k} D_m^l(u)\alpha_i(l,r)D_{k-m}^r(u) + \sum_{s=1}^{d} B(s,i)D_k^s(u)t^k\right)t^k - \gamma_i \\
&= \sum_{k=0}^{\infty}\left(\frac{1}{2}\sum_{m=0}^{k} D_m^\top(u)\alpha_i D_{k-m}(u) + \beta_i^T D_k(u)\right)t^k - \gamma_i \\
&= (\frac{1}{2}u^\top\alpha_i u + \beta_i^\top u - \gamma_i) + \sum_{k=1}^{\infty}\left(\frac{1}{2}\sum_{m=0}^{k} D_m^\top(u)\alpha_i D_{k-m}(u) + \beta_i^\top D_k(u)\right)t^k.
\end{aligned}
\tag{A.5}
$$

Comparing the Taylor coefficients in (A.4) and (A.5), we obtain

$$
\begin{aligned}
D_1(u) &= \frac{1}{2}u^\top\alpha_i u + \beta_i^\top u - \gamma_i, \\
D_{k+1}^i(u) &= \frac{1}{1+k}\left(\frac{1}{2}\sum_{n=0}^{k} D_n^\top(u)\alpha_i D_{k-n}(u) + \beta_i^\top D_k(u)\right).
\end{aligned}
$$

Similarly, we also obtain

$$
\begin{aligned}
C_1(u) &= \frac{1}{2}u^\top au + b^\top u - c, \\
C_{k+1}(u) &= \frac{1}{1+k}\left(\frac{1}{2}\sum_{n=0}^{k} D_n^\top(u)aD_{k-n}(u) + b^\top D_k u\right).
\end{aligned}
$$

□

This proposition allows us to approximate the $(\phi(t,u), \psi(t,u))$ by

$$
\begin{aligned}
\phi(t,u) &\approx \sum_{k=0}^{N} C_k(u)t^k, \\
\psi_i(t,u) &\approx \sum_{k=0}^{N} D_k^i(u)t^k.
\end{aligned}
$$

The approximation errors are of the form $\sum_{k=N+1}^{\infty} A_k(u)t^k$. The approximation is accurate and converges quickly if $t \approx 0$. For $t \gg 0$, we divide the time interval into several subintervals which are small enough to make the approximation accurate.

Choose time steps $\Delta_i > 0, i = 1, \cdots, n$ such that $T - t = \Delta_1 + \cdots + \Delta_n$, then by the tower property,

$$
e^{\phi(T-t,u)+\psi(T-t,u)^\top X(t)} = \mathbb{E}[e^{u^\top X(T)} \mid \mathcal{F}_t]
$$

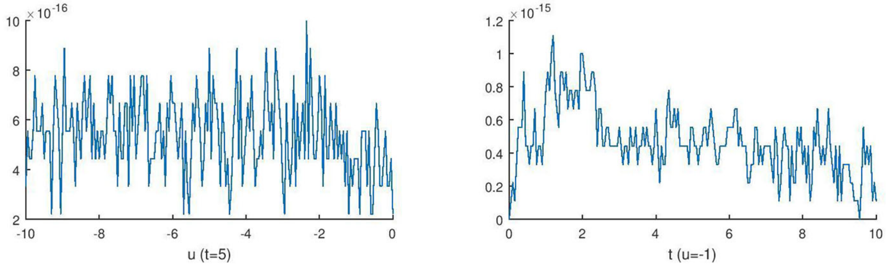

Fig. 12 Numerical errors of the approximate solutions

$$
\begin{aligned}
&= \mathbb{E}[\mathbb{E}[e^{u^\top X(T)} \mid \mathcal{F}_{T-\Delta_1}] \mid \mathcal{F}_t] \\
&= e^{\phi(\Delta_1,u)}\mathbb{E}[e^{\psi(\Delta_1,u)^\top X(T-\Delta_1)} \mid \mathcal{F}_t] \\
&= e^{\phi(\Delta_1,u)}\mathbb{E}[\mathbb{E}[e^{\psi(\Delta_1,u)^\top X(T-\Delta_1)} \mid \mathcal{F}_{T-\Delta_1-\Delta_2}] \mid \mathcal{F}_t] \\
&= e^{\phi(\Delta_1,u)+\phi(\Delta_2,\psi(\Delta_1,u))}\mathbb{E}[e^{\psi(\Delta_2,\psi(\Delta_1,u))^\top X(T-\Delta_1-\Delta_2)} \mid \mathcal{F}_t] \\
&\vdots \\
&= e^{\phi(\Delta_1,u_0)+\phi(\Delta_2,u_1)+\cdots+\phi(\Delta_n,u_{n-1})}e^{\psi(\Delta_n,u_{n-1})^\top X(t)},
\end{aligned}
$$

where $u_{i+1} = \psi(\Delta_{i+1}, u_i)$, $u_0 = u$.

Comparing the two extreme sides of the latter equation, we obtain

$$
\begin{aligned}
\phi(T-t,u) &= \sum_{i=1}^{n} \phi(\Delta_i, u_{i-1}), \\
\psi(T-t,u) &= \psi(\Delta_n, u_{n-1}).
\end{aligned}
$$

In practice, we can set the approximation error level to be ε. If at each step we choose $\Delta = (\varepsilon A_N(u))^{\frac{1}{N}}$, then the last term in the Taylor expansion is $A_N(u) * \Delta^N = \varepsilon$. Hence we can control the approximation at the level ε.

Remark A.1 The values of the functions $\phi(t,u)$, $\psi(t,u)$ might go to infinity for some value of t and u . In these situations, the Taylor expansion approximation doesn't work. However, in financial application, we assume these cases do not exist since in finance we always assume the moments of the underlying process exist.

Example A.1 Consider (A.3) with $d = 1$ and the admissible parameters $\alpha = \gamma = 1$, $\beta = -1$. The ODE for ψ is

$$
\begin{aligned}
\partial_t \psi(t,u) &= \psi(t,u)^2 - \psi(t,u) - 1, \\
\psi(0,u) &= u.
\end{aligned}
$$

The unique closed form solution to this equation is given by, see (Filipović 2009, Eq.(10.47)),

$$
\psi(t,u) = \frac{2(e^{\sqrt{5}t} - 1) - ((\sqrt{5}-1)e^{\sqrt{5}t} + \sqrt{5} + 1)u}{(\sqrt{5}+1)e^{\sqrt{5}t} + \sqrt{5} - 1 - 2(e^{\sqrt{5}t} - 1)u}.
$$

For the numerical approximation, we choose the Taylor expansion order $N = 10$ and the tolerance of the error $\epsilon = 10^{-16}$. The plots in Figure 12 show the numerical errors for different t and u.

References

Andrieu C, Doucet A (2002) Particle filtering for partially observed Gaussian state space models. J R Stat Soc Ser B 64(4):827–836. https://doi.org/10.1111/1467-9868.00363

Arulampalam M, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans Signal Process 50(2):174–188

Bai J, Perron P (1998) Estimating and testing linear models with multiple structural changes. Econometrica 66(1), 47–78. http://www.jstor.org/stable/2998540

Bruno M (2013) Sequential Monte Carlo methods for nonlinear discrete-time filtering. Synth Lect Signal Process 6(1):1–99

Cappé O, Godsill S, Moulines E (2007) An overview of existing methods and recent advances in sequential Monte Carlo. Proc IEEE 95(5):899–924

Chen R, Liu J (2000) Mixture Kalman filter. J R Stat Soc Ser B 62:493–508. https://doi.org/10.1111/1467-9868.00246

Chib S (1998) Estimation and comparison of multiple change-point models. J Econ 86(2):221–241

Chopin N (2002) A sequential particle filter method for static models. Biometrika 89(3):539–551. https://doi.org/10.1093/biomet/89.3.539

Chopin N (2007) Dynamic detection of change points in long time series. Ann Inst Stat Math 59(2):349–366

Chopin N, Jacob P, Papaspiliopoulos O (2013) SMC2: an efficient algorithm for sequential analysis of state space models. J R Stat Soc Ser B (Methodological) 75(3):397–426. https://doi.org/10.1111/j.1467-9868.2012.01046.x

Chui C, Chen G (2017) Kalman filtering with real-time applications, 5th edn. Springer, Cham. https://doi.org/10.1007/978-3-319-47612-4

Cox J, Ingersoll J, Ross S (1985) A theory of the term structure of interest rates. Econometrica 53(2):385–407. https://doi.org/10.2307/1911242

Crisan D, Míguez J (2018) Nested particle filters for online parameter estimation in discrete-time state-space Markov models. Bernoulli 24(4A):3039–3086. https://doi.org/10.3150/17-BEJ954

Doucet A (1997) Monte carlo methods for bayesian estimation of hidden markov models. application to radiation signals. Ph.D. thesis, Univ. Paris-Sud, Orsay

Doucet A, Godsill S, Andrieu C (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. Stat Comput 10(3):197–208

Doucet A, Johansen A (2009) A tutorial on particle filtering and smoothing: fifteen years later. Handbook of nonlinear filtering 12(656–704):3

Duffie D, Filipović D, Schachermayer W (2003) Affine processes and applications in finance. Ann Appl Probabil 13(3):984–1053

Duffie D, Pan J, Singleton K (2000) Transform analysis and asset pricing for affine jump-diffusions. Econometrica 68(6):1343–1376

Einicke G, White L (1999) Robust extended Kalman filtering. IEEE Trans Signal Process 47(9):2596–2599

Fearnhead P, Liu Z (2007) On-line inference for multiple change-point problems. J R Stat Soc Ser B (Stat Methodol) 69(4):589–605

Filipović D (2009) Term-Structure Models: a Graduate Course. Springer. https://doi.org/10.1007/978-3-540-68015-4

Gamerman D, Lopes H (2006) Markov chain Monte Carlo, second edn. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL. Stochastic simulation for Bayesian inference

Geweke J, Tanizaki H (1999) On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state-space models. Commun Stat Simul Comput 28(4):867–894

Glasserman P (2013) Monte Carlo methods in financial engineering, vol 53. Springer, Berlin

Gordon N, Salmond D, Smith A (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In: IEE proceedings F (radar and signal processing), vol 140, pp 107–113. IET

Grewal M, Andrews A (2015) Kalman filtering, 4th edn. Theory and practice using MATLAB$textregister$, Wiley, Hoboken, NJ

He Z, Maheu J (2010) Real time detection of structural breaks in GARCH models. Comput Stat Data Anal 54(11):2628–2640

Heston S (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev Financ Stud 6(2):327–343

Higdon D (1998) Auxiliary variable methods for Markov chain Monte Carlo with applications. J Am Stat Assoc 93(442):585–595

Hull J, White A (1990) Pricing interest-rate-derivative securities. Rev Financ Stud 3(4):573–592

Kantas N, Doucet A, Singh S, Maciejowski J, Chopin N (2015) On particle methods for parameter estimation in state-space models. Statistical Science. A Review J Inst Math Stat 30(3):328–351. https://doi.org/10.1214/14-STS511

Karatzas I, Shreve S (1998) Brownian motion. Springer, Berlin

Keller-Ressel M, Mayerhofer E (2015) Exponential moments of affine processes. Ann Appl Probabil 25:151–168

Kitagawa G (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. J Comput Graph Stat 5(1):1–25. https://doi.org/10.2307/1390750

Liu J, Chen R (1998) Sequential Monte Carlo methods for dynamic systems. J Am Stat Assoc 93(443):1032–1044. https://doi.org/10.2307/2669847

Liu J, West M (2001) Combined parameter and state estimation in simulation-based filtering. In: Sequential Monte Carlo methods in practice, Stat. Eng. Inf. Sci., pp. 197–223. Springer, New York

Ljung L, Söderström T (1983) Theory and practice of recursive identification. MIT Press, New York

Maheu J, Gordon S (2008) Learning, forecasting and structural breaks. J Appl Econ 23(5):553–583

Miguez J, Bugallo M, Djuric P (2005) Novel particle filtering algorithms for fixed parameter estimation in dynamic systems. In Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA) pp. 46 – 51. https://doi.org/10.1109/ISPA.2005.195382

Nemeth C, Fearnhead P, Mihaylova L (2014) Sequential Monte Carlo methods for state and parameter estimation in abruptly changing environments. IEEE Trans Signal Process 62(5):1245–1255. https://doi.org/10.1109/TSP.2013.2296278

Papavasiliou A (2006) Parameter estimation and asymptotic stability in stochastic filtering. Stochast Processes Appl 116(7):1048–1065. https://doi.org/10.1016/j.spa.2006.01.002

Pérez-Vieites S, Mariño IP, Míguez J (2018) Probabilistic scheme for joint parameter estimation and state prediction in complex dynamical systems. Phys Rev E 98:063305. https://doi.org/10.1103/PhysRevE.98.063305

Pitt M, Shephard N (1999) Filtering via simulation: auxiliary particle filters. J Am Stat Assoc 94(446):590–599. https://doi.org/10.2307/2670179

Press, S.: Subjective and objective Bayesian statistics, Second edn. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ (2003) Principles, models, and applications. With contributions by Siddhartha Chib, Merlise Clyde, George Woodworth and Alan Zaslavsky

Robert C (2007) The Bayesian choice, second edn. Springer Texts in Statistics. Springer, New York. From decision-theoretic foundations to computational implementation

van Schuppen J (1989) Stochastic realization problems. In: Three decades of mathematical system theory, Lect. Notes Control Inf. Sci., vol. 135, pp. 480–523. Springer, Berlin. https://doi.org/10.1007/BFb0008474

Sorenson H, Alspach D (1971) Recursive Bayesian estimation using Gaussian sums. Automatica 7(4):465–479

Stroud J, Katzfuss M, Wikle C (2018) A Bayesian adaptive ensemble Kalman filter for sequential state and parameter estimation. Monthly Weather Rev 146(1):373–386

Wan E, van der Merwe R (2002) The Unscented Kalman Filter, chap. 7, pp. 221–280. Wiley. https://doi.org/10.1002/0471221546.ch7

Wan E, Nelson A (2002) Dual Extended Kalman Filter Methods, chap. 5, pp. 123–173. Wiley. https://doi.org/10.1002/0471221546.ch5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.