ORIGINAL PAPER

Statistics for Stochastic Processes

Nonparametric Bayesian volatility learning under microstructure noise

Shota Gugushvili[1] [ID] · **Frank van der Meulen[2]** [ID] · **Moritz Schauer[3]** [ID] ·
Peter Spreij[4,5] [ID]

Abstract

In this work, we study the problem of learning the volatility under market microstructure noise. Specifically, we consider noisy discrete time observations from a stochastic differential equation and develop a novel computational method to learn the diffusion coefficient of the equation. We take a nonparametric Bayesian approach, where we a priori model the volatility function as piecewise constant. Its prior is specified via the inverse Gamma Markov chain. Sampling from the posterior is accomplished by incorporating the Forward Filtering Backward Simulation algorithm in the Gibbs sampler. Good performance of the method is demonstrated on two representative synthetic data examples. We also apply the method on a EUR/USD exchange rate dataset. Finally we present a limit result on the prior distribution.

✉ Frank van der Meulen
    f.h.van.der.meulen@vu.nl

    Shota Gugushvili
    shota@yesdatasolutions.com

    Moritz Schauer
    smoritz@chalmers.se

    Peter Spreij
    spreij@uva.nl

1   Biometris, Wageningen University & Research, Postbus 16, 6700 AA Wageningen, The
    Netherlands

2   Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV
    Amsterdam, The Netherlands

3   Department of Mathematical Sciences, Chalmers University of Technology and University of
    Gothenburg, 412 96 Gothenburg, Sweden

4   Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090
    GE Amsterdam, The Netherlands

5   Institute for Mathematics, Astrophysics and Particle Physics, Radboud University, Nijmegen, The
    Netherlands

        ⌂ Springer

1 Introduction

Let the one-dimensional stochastic differential equation (SDE)

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + s(t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T], \tag{1}$$

be given. Here W is a standard Wiener process and b and s are referred to as the drift function and volatility function, respectively. We assume (not necessarily uniformly spaced) observation times $\{t_i : 1 \le i \le n\}$ and observations $\mathscr{Y}_n = \{Y_1, \ldots, Y_n\}$, where

$$Y_i = X_{t_i} + V_i, \quad 0 < t_1 < \cdots < t_n = T, \tag{2}$$

and $\{V_i\}$ is a sequence of independent and identically distributed random variables, independent of W. Our aim is to learn the volatility s using the noisy observations $\mathscr{Y}_n$. Knowledge of the volatility is of paramount importance in financial applications, specifically in pricing financial derivatives, see, e.g., Musiela and Rutkowski (2005), and in risk management.

The quantity $\Delta t_i = t_i - t_{i-1}$ is referred to as the observation density, especially if the time instants are equidistant, and then $1/\Delta t_i$ will be the frequency of the observations. Small values of Δt_i correspond to high frequency, dense-in-time data. Intraday financial data are commonly thought to be high frequency data. In this high frequency financial data setting, which is the one we are interested in the present work, the measurement errors $\{V_i\}$ are referred to as microstructure noise. Their inclusion in the model aims to reflect such features of observed financial time series as their discreteness or approximations due to market friction. Whereas for low-frequency financial data these can typically be neglected without much ensuing harm, empirical evidence shows that this is not the case for high frequency data; cf. Mykland and Zhang (2012).

There exists a large body of statistics and econometrics literature on nonparametric volatility estimation under microstructure noise. See, e.g., Hoffmann et al. (2012), Jacod et al. (2009), Mykland and Zhang (2009), Reiß (2011), Sabel et al. (2015); a recent overview is Mykland and Zhang (2012). The literature predominantly deals with estimation of the integrated volatility $\int_0^t s^2(u)\mathrm{d}u$, although inference on s has also been studied in several of these references. Various methods proposed in the above-mentioned works share the property of being *frequentist* in nature. An important paper (Mancini et al., 2015) is a mix of theoretical and practical results, with a clear predominance of the former. Its main purpose (see page 262) is proposing a unifying frequentist approach to spot volatility estimation, from which many previously existing approaches would be derived as special cases. This allows comparison between them.

As our approach is Bayesian, our results do not fall under the general umbrella of Mancini et al. (2015).

In this paper, our main and novel contribution is the development of a *practical* nonparametric Bayesian approach to volatility learning under microstructure noise. We specify an inverse Gamma Markov chain prior (Cf. Cemgil & Dikmen 2007) on the volatility function *s* and reduce our model to the Gaussian linear state space model. Posterior inference in the latter is performed using Gibbs sampling including a Forward Filtering Backward Simulation (FFBS) step. We demonstrate good performance of our method on two representative simulation examples. The first example uses a benchmark function, popular in nonparametric regression, see Fan and Gijbels (1995), as the volatility function. In the second example we consider a well known and widely used stochastic volatility model, the Heston model, see Heston (1993), or (Filipovic 2009, Section 10.3.3) and (Brigo & Mercurio 2006, Chapter 19, Appendix A). In both examples our approach shows accurate results. We also apply our method to a real data set of EUR/USD exchange rates and we deduce a clear and understandable variation in the volatility over time.

In general, a nonparametric approach reduces the risk of model misspecification; the latter may lead to distorted inferential conclusions. The presented nonparametric method is not only useful for an exploratory analysis of the problem at hand (cf. Silverman 1986), but also allows honest representation of inferential uncertainties (cf. Müller & Mitra 2013). Attractive features of the Bayesian approach include its internal coherence, automatic uncertainty quantification in parameter estimates via Bayesian credible sets, and the fact that it is a fundamentally likelihood-based method. For a modern monographic treatment of nonparametric Bayesian statistics see Ghosal and van der Vaart (2017); an applied perspective is found in Müller et al. (2015).

The paper is organised as follows: in Sect. 2 we introduce in detail our approach, followed by Sect. 3 where the limiting behaviour of the prior on the squared volatility is derived under mesh refinement. In Sect. 4 we test its practical performance on synthetic data examples. Section 5 applies our method on a real data example. Section 6 summarises our findings. Finally, Appendix A gives further implementational details.

Notation

We denote the inverse Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ by IG(α, β). Its density is

$$x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}, \quad x > 0.$$

By $N(\mu, \sigma^2)$ we denote a normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. The uniform distribution on an interval $[a, b]$ is denoted by Uniform(a, b). For a random variate X, the notation $X \sim p$ stands for the fact that X is distributed according to a density p, or is drawn according to a density p. Conditioning of a random variate X on a random variate Y is denoted by $X \mid Y$. By $\lfloor x \rfloor$ we denote the integer part of a real number x. The notation $p \propto q$ for a density p denotes the fact that a positive

function q is an unnormalised density corresponding to p: p can be recovered from q as $q / \int q$. Finally, we use the shorthand notation $a_{k:\ell} = (a_k, \dots, a_\ell)$.

2 Methodology

In this section we introduce our methodology for inferring the volatility. We first recast the model into a simpler form that is amenable to computational analysis, next specify a nonparametric prior on the volatility, and finally describe an MCMC method for sampling from the posterior.

2.1 Linear state space model

Let $t_0 = 0$. By Eq. (1), we have

$$X_{t_i} = X_{t_{i-1}} + \int_{t_{i-1}}^{t_i} b(t, X_t) \mathrm{d}t + \int_{t_{i-1}}^{t_i} s(t) \mathrm{d}W_t. \tag{3}$$

We derive our method under the assumption that the "true", data-generating volatility s is a deterministic function of time t. Next, if the "true" s is in fact a stochastic process, we apply our procedure without further changes, as if s were deterministic. As shown in the example of Sect. 4.2, this works in practice. That this is the case is easiest to understand in the situation where one can discern a two-stage procedure. First the stochastic volatility is generated, and given a realization of it, the observations are generated by an independent Brownian motion W. In Kanaya and Kristensen (2016) such an approach is used for simulation.

Over short time intervals $[t_{i-1}, t_i]$, the term $\int_{t_{i-1}}^{t_i} s(t) \mathrm{d}W_t$ in (3), roughly speaking, will dominate the term $\int_{t_{i-1}}^{t_i} b(t, X_t) \mathrm{d}t$, as the latter scales as Δt_i, whereas the former as $\sqrt{\Delta t_i}$ (due to the properties of the Wiener process paths). As our emphasis is on learning s rather than b, following (Gugushvili et al., 2020, 2019b) we act as if the process X had a zero drift, $b \equiv 0$. The justification of this procedure is explained (Gugushvili et al., 2020). A similar idea is often used in *frequentist* volatility estimation procedures in the high frequency financial data setting; see Mykland and Zhang (2012), Sect. 2.1.5 for an intuitive exposition. Formal results why this works in specific settings rely on Girsanov's theorem, see, e.g., Gugushvili et al. (2020), Hoffmann et al. (2012), Mykland and Zhang (2009). Further reasons why one would like to set $b = 0$ are that b is a *nuisance parameter*, in specific applications its appropriate parametric form might be unknown, and finally, a single observed time series is not sufficient to learn b consistently (see Ignatieva & Platen 2012).

We thus *assume* $X_{t_i} = X_{t_{i-1}} + U_i$, where $U_i = \int_{t_{i-1}}^{t_i} s(t) \mathrm{d}W_t$. Note that then

$$U_i \sim N(0, w_i) \quad \text{with} \quad w_i = \int_{t_{i-1}}^{t_i} s^2(t) \mathrm{d}t, \tag{4}$$

and also that $\{U_i\}$ is a sequence of independent random variables. To simplify our notation, write $x_i = X_{t_i}$, $y_i = Y_i$, $u_i = U_i$, $v_i = V_i$. The preceding arguments and (2) allow us to reduce our model to the *linear state space model*

$$\begin{aligned} x_i &= x_{i-1} + u_i, \\ y_i &= x_i + v_i, \end{aligned} \qquad (5)$$

where $i = 1, \ldots, n$. The first equation in (5) is the *state equation*, while the second equation is the *observation equation*. We assume that $\{v_i\}$ is a sequence of independent $N(0, \eta_v)$ distributed random variables, independent of the Wiener process W in (1), so that $\{v_i\}$ is independent of $\{u_i\}$. For justification of such assumptions on the noise sequence $\{v_i\}$ from a practical point of view, see Sabel et al. (2015), page 229. We endow the *initial state* x_0 with the $N(\mu_0, C_0)$ prior distribution. Then (5) is a *Gaussian* linear state space model. This is very convenient computationally. Had we not followed this route, we would have had to deal with an *intractable likelihood*, which constitutes the main *computational bottleneck* for Bayesian inference in SDE models; see, e.g, Papaspiliopoulos et al. (2013) and van der Meulen and Schauer (2017) for discussion.

2.2 Prior

For the measurement error variance η_v, we assume a priori $\eta_v \sim \text{IG}(\alpha_v, \beta_v)$. The construction of the prior for s is more complex and follows (Gugushvili et al., 2019b), that in turn relies on Cemgil and Dikmen (2007). Fix an integer $m < n$. Then we have a unique decomposition $n = mN + r$ with $0 \le r < m$, where $N = \lfloor n/m \rfloor$. Now define bins $B_k = [t_{m(k-1)}, t_{mk})$, $k = 1, \ldots, N-1$, and $B_N = [t_{m(N-1)}, T]$. We model s as

$$s = \sum_{k=1}^{N} \xi_k \mathbf{1}_{B_k}, \qquad (6)$$

where N (the number of bins) is a hyperparameter. Then $s^2 = \sum_{k=1}^{N} \theta_k \mathbf{1}_{B_k}$, where $\theta_k = \xi_k^2$. We complete the prior specification for s by assigning a prior distribution to the coefficients $\theta_{1:N}$. For this purpose, we introduce auxiliary variables $\zeta_{2:N}$, and suppose the sequence $\theta_1, \zeta_2, \theta_2, \ldots, \zeta_k, \theta_k, \ldots, \zeta_N, \theta_N$ forms a Markov chain (in this order of variables). The transition distributions of the chain are defined by

$$\theta_1 \sim \text{IG}(\alpha_1, \beta_1), \quad \zeta_{k+1}|\theta_k \sim \text{IG}(\alpha, \alpha\theta_k^{-1}), \quad \theta_{k+1}|\zeta_{k+1} \sim \text{IG}(\alpha, \alpha\zeta_{k+1}^{-1}), \qquad (7)$$

where $\alpha_1, \beta_1, \alpha$ are hyperparameters. We refer to this chain as an inverse Gamma Markov chain, see Cemgil and Dikmen (2007). The corresponding prior on $\theta_{1:N}$ will be called the inverse Gamma Markov chain (IGMC) prior. The definition in (7) ensures that $\theta_1, \ldots, \theta_N$ are positively correlated, which imposes *smoothing* across different bins B_k. Simultaneously, it ensures partial conjugacy in the Gibbs sampler that we derive below, leading to simple and tractable MCMC inference. In our experience, an uninformative choice $\alpha_1, \beta_1 \to 0$ performs well in practice. We also endow α with a prior distribution and assume $\log \alpha \sim N(a, b)$, with hyperparameters $a \in \mathbb{R}, b > 0$

chosen so as to render the hyperprior on α diffuse. As explained in Gugushvili et al. (2020), Gugushvili et al. (2019b), the hyperparameter N (or equivalently m) can be considered both as a smoothing parameter and the resolution at which one wants to learn the volatility function. Obviously, given the limited amount of data, this resolution cannot be made arbitrarily fine. On the other hand, as shown in Gugushvili et al. (2019b) (see also Gugushvili et al. 2019a), inference with the IGMC prior is quite robust with respect to a wide range of values of N, as the corresponding Bayesian procedure has an additional *regularisation parameter* α that is learned from the data. Statistical optimality results in Munk and Schmidt-Hieber (2010) suggest that in our setting N should be chosen considerably smaller than in the case of an SDE observed without noise (that was studied via the IGMC prior in Gugushvili et al. (2019b)).

2.3 Likelihood

Although an expression for the posterior of s can be written down in closed form, it is not amenable to computations. This problem is alleviated by following a data augmentation approach, in which $x_{0:n}$ are treated as *missing data*, whereas the $y_{1:n}$ are the observed data; cf. Tanner and Wong (1987). An expression for the joint density of all random quantities involved is easily derived from the prior specification and (5). We have

$$
\begin{aligned}
&p(y_{1:n}, x_{0:n}, \theta_{1:N}, \zeta_{2:N}, \alpha, \eta_v) \\
&= \left(\prod_{k=1}^{n} p(y_k \mid x_k, \eta_v) \right) p(x_{0:n} \mid \theta_{1:N}) \\
&\quad \times p(\theta_1) \prod_{k=1}^{N-1} \left[p(\zeta_{k+1} \mid \theta_k, \alpha) p(\theta_{k+1} \mid \zeta_{k+1}, \alpha) \right] p(\alpha) p(\eta_v).
\end{aligned}
$$

Except for $p(x_{0:n} \mid \theta_{1:N})$, all the densities have been specified directly in the previous subsections. To obtain an expression for the latter, define (with $\Delta_i \equiv \Delta t_i$)

$$
Z_k = \sum_{i=(k-1)m+1}^{km} \frac{(x_i - x_{i-1})^2}{\Delta_i}, \quad k = 1, \ldots, N-1,
$$

$$
Z_N = \sum_{i=(N-1)m+1}^{n} \frac{(x_i - x_{i-1})^2}{\Delta_i},
$$

and set $m_k = m$ for $k = 1, \ldots, N-1$, and $m_N = m + r$. Then

$$
p(x_{0:N} \mid \theta_{1:N}) \propto e^{(x_0 - \mu_0)^2 / (2C_0)} \prod_{k=1}^{N} \theta_k^{-m_k/2} \exp\left(-\frac{Z_k}{2\theta_k} \right).
$$

2.4 Gibbs sampler

We use the Gibbs sampler to sample from the joint conditional distribution of $(x_{0:n}, \theta_{1:N}, \zeta_{2:N}, \eta_v, \alpha)$ given $y_{1:n}$. The full conditionals of $\theta_{1:N}, \zeta_{2:N}, \eta_v$ are easily derived from Sect. 2.3 and recognised to be of the inverse Gamma type, see Sect. A.2. The parameter α can be updated via a Metropolis-Hastings step. For updating $x_{0:N}$, conditional on all other parameters, we use the standard Forward Filtering Backward Simulation (FFBS) algorithm for Gaussian state space models (cf. Section 4.4.3 in Petris et al. 2009). The resulting Gibbs sampler is summarised in Algorithm 1. For details, see Appendix A.

Algorithm 1: Gibbs sampler for volatility learning

Data: Observations $y_{1:n}$
Hyperparameters α_1, β_1, α_v, β_v, a, b, N;
Result: Posterior samples $\theta_{1:N}^i : i = 1, \ldots, M$
Initialization $\theta_{1:N}^0, \zeta_{1:N}^0, \eta_v^0, \alpha^0$;
while $i \leq M$ **do**
    sample $x_{0:n}^i$ via FFBS;
    sample $\theta_{1:N}^i$ from the inverse Gamma full conditionals;
    sample $\zeta_{2:N}^i$ from the inverse Gamma full conditionals;
    sample η_v^i from the inverse Gamma full conditional;
    sample α^i via a Metropolis-Hastings step;
    set $i = i + 1$.
end

---

3 Asymptotics for the prior on the squared volatility

We first provide in Proposition 1 results on the prior conditional mean and variance of the θ_k as in Sect. 2.2. These results will be exploited to find an asymptotic regime for the θ_k when the number of bins tends to infinity. In this section we depart from the original setting with α as a random hyperparameter (having a lognormal distribution), but instead we take it as a deterministic one that we let grow to infinity to obtain our asymptotic results.

Proposition 1 *If $\alpha > 2$ is a fixed parameter then the IGMC prior of* Sect. 2.2 *satisfies*

$$\mathbb{E}_k[\theta_{k+1} - \theta_k] = \frac{1}{\alpha - 1}\theta_k, \tag{8}$$

$$\mathbb{V}_k(\theta_{k+1} - \theta_k) = \frac{\alpha(2\alpha - 1)}{(\alpha - 1)^2(\alpha - 2)}\theta_k^2, \tag{9}$$

where $\mathbb{E}_k$ and $\mathbb{V}_k$ respectively denote expectation and variance, conditional on θ_k. Consequently, the conditional mean squared error $\mathbb{E}_k(\theta_{k+1} - \theta_k)^2$ equals $\frac{2(\alpha+1)\theta_k^2}{(\alpha-1)(\alpha-2)}$.

Proof We will use that for $Z \sim \text{IG}(a, b)$ it holds that

$$\mathbb{E}Z^{-1} = \frac{a}{b},$$
$$\mathbb{E}Z^{-2} = \frac{a(a+1)}{b^2}. \tag{10}$$

First we consider the conditional mean. Since $\theta_{k+1} \mid \zeta_{k+1} \sim \text{IG}(\alpha, \alpha\zeta_{k+1}^{-1})$ we have

$$\mathbb{E}[\theta_{k+1} \mid \zeta_{k+1}] = \frac{\alpha\zeta_{k+1}^{-1}}{\alpha - 1} \tag{11}$$

provided $\alpha > 1$. Exploiting that the sequence $\theta_1, \zeta_2, \theta_2, \ldots, \zeta_k, \theta_k, \ldots, \zeta_N, \theta_N$ forms a Markov chain (in this order of variables), one has

$$\mathbb{E}_k\theta_{k+1} = \mathbb{E}_k\mathbb{E}[\theta_{k+1} \mid \theta_k, \zeta_{k+1}] = \mathbb{E}_k\mathbb{E}[\theta_{k+1} \mid \zeta_{k+1}] = \frac{\alpha}{\alpha - 1}\mathbb{E}_k\zeta_{k+1}^{-1},$$

where we used (11) at the last equality sign. Using (10) and $\zeta_{k+1} \mid \theta_k \sim \text{IG}(\alpha, \alpha\theta_k^{-1})$ one obtains

$$\mathbb{E}_k\theta_{k+1} = \frac{\alpha}{\alpha - 1}\frac{\alpha}{\alpha\theta_k^{-1}} = \frac{\alpha}{\alpha - 1}\theta_k,$$

which is equivalent to (8).

Next we calculate the conditional variance. We have

$$\mathbb{V}[\theta_{k+1} \mid \zeta_{k+1}] = \frac{\alpha^2\zeta_{k+1}^{-2}}{(\alpha - 1)^2(\alpha - 2)} \tag{12}$$

provided $\alpha > 2$. We need the following variation on the law of total variance. If $X \in L^2(\Omega, \mathscr{F}, \mathbb{P})$, and $\mathscr{G}, \mathscr{H}$ are subsigma-algebras of $\mathscr{F}$ with $\mathscr{H} \subset \mathscr{G}$. Then, with $X^{\mathscr{G}} = \mathbb{E}[X \mid \mathscr{G}]$, it holds that

$$\text{Var}(X \mid \mathscr{H}) = \mathbb{E}[\text{Var}(X \mid \mathscr{G}) \mid \mathscr{H}] + \text{Var}(X^{\mathscr{G}} \mid \mathscr{H}).$$

We use this result with $X = \theta_{k+1}, \mathscr{G} = \sigma(\theta_k, \zeta_{k+1}), \mathscr{H} = \sigma(\theta_k)$, obtaining

$$\text{Var}(\theta_{k+1} \mid \theta_k) = \mathbb{E}[\text{Var}(\theta_{k+1} \mid \theta_k, \zeta_{k+1}) \mid \theta_k] + \text{Var}(\mathbb{E}[\theta_{k+1} \mid \theta_k, \zeta_{k+1}] \mid \theta_k)$$
$$= \mathbb{E}[\text{Var}(\theta_{k+1} \mid \zeta_{k+1}) \mid \theta_k] + \text{Var}(\mathbb{E}[\theta_{k+1} \mid \zeta_{k+1}] \mid \theta_k),$$

as now $X^{\mathscr{G}} = \mathbb{E}[\theta_{k+1} \mid \theta_k, \zeta_{k+1}] = \mathbb{E}[\theta_{k+1} \mid \zeta_{k+1}]$ and $\text{Var}(\theta_{k+1} \mid \theta_k, \zeta_{k+1}) = \text{Var}(\theta_{k+1} \mid \zeta_{k+1})$ in view of the Markov property. Hence, in our abbreviated notation,

$$\mathbb{V}_k\theta_{k+1} = \mathbb{E}_k\mathbb{V}(\theta_{k+1} \mid \zeta_{k+1}) + \mathbb{V}_k\mathbb{E}[\theta_{k+1} \mid \zeta_{k+1}].$$

Hence, using (11) and (12)

$$
\begin{aligned}
\mathbb{V}_k \theta_{k+1} &= \frac{\alpha^2}{(\alpha-1)^2(\alpha-2)} \mathbb{E}_k \zeta_{k+1}^{-2} + \mathbb{V}_k \frac{\alpha \zeta_{k+1}^{-1}}{\alpha-1} \\
&= \left(\frac{\alpha^2}{(\alpha-1)^2(\alpha-2)} + \frac{\alpha^2}{(\alpha-1)^2} \right) \mathbb{E}_k \zeta_{k+1}^{-2} - \frac{\alpha^2}{(\alpha-1)^2} \left(\mathbb{E}_k \zeta_{k+1}^{-1} \right)^2 \\
&= \frac{\alpha^2}{(\alpha-1)(\alpha-2)} \mathbb{E}_k \zeta_{k+1}^{-2} - \frac{\alpha^2}{(\alpha-1)^2} \left(\mathbb{E}_k \zeta_{k+1}^{-1} \right)^2 \\
&= \frac{\alpha^2}{(\alpha-1)(\alpha-2)} \frac{\alpha(\alpha+1)}{\alpha^2 \theta_k^{-2}} - \frac{\alpha^2}{(\alpha-1)^2} \left(\frac{\alpha}{\alpha \theta_k^{-1}} \right)^2 \\
&= \frac{\alpha(2\alpha-1)}{(\alpha-1)^2(\alpha-2)} \theta_k^2,
\end{aligned}
$$

which establishes (9). The expression for the mean squared error follows from (8) and (9). □

The expression for the conditional mean squared error in Proposition 1 shows that it decreases in α, which illustrates the regularising property of this parameter. Therefore we are interested in the behaviour of the prior distribution on the θ_k for large values of α. We will scale α with the number of bins N to obtain a limit result by applying Donsker's theorem (Cf. Corollary VII.3.11 in Jacod & Shiryaev 2013), upon letting $N \to \infty$. We give some heuristics to derive the limit behaviour.

Take $\alpha = \gamma N$, where in the latter expression γ is a positive scaling factor. As the law of the θ_k depends on γN, we are thus interested in the law of θ_k for $N \to \infty$. This entails simultaneously increasing the number of bins as also the dependence of the values on the bins. Below we argue that under the IGMC-prior, with θ_1 fixed, the process $t \mapsto s^2(t)$, with $s^2(t) = \sum_{k=1}^N \theta_k \mathbf{1}_{B_k}(t)$, converges weakly to the continuous time process $t \mapsto \theta_1 Z_t$ where

$$
\log Z_t = \sqrt{\frac{2}{\gamma}} W_t - \frac{1}{\gamma} t. \tag{13}
$$

The expressions for conditional mean and variance of Proposition 1 are for large N (in particular then $\alpha = \gamma N > 2$) approximately equal to

$$
\mathbb{E}_k \theta_{k+1} = \theta_k, \tag{14}
$$

$$
\mathbb{V}_k(\theta_{k+1}) = \frac{2\theta_k^2}{\gamma N}. \tag{15}
$$

These properties are shared with θ_k generated by the recursion

$$
\theta_{k+1} = \theta_k \left(1 + \sqrt{\frac{2}{\gamma N}} z_k \right),
$$

where, for all k, z_k is independent of $\theta_1, \ldots, \theta_k$, $\mathbb{E}z_k = 0$, $\mathbb{V}\text{ar }z_k = 1$. The z_k can be seen as martingale differences, or even as IID random variables. Solving the recursion starting from θ_1 after taking logarithms leads to

$$
\begin{aligned}
\log \theta_{k+1} - \log \theta_1 &= \sum_{j=1}^{k} \log\left(1 + \sqrt{\frac{2}{\gamma N}}z_j\right) \\
&\approx \sum_{j=1}^{k} \sqrt{\frac{2}{\gamma N}}z_j - \sum_{j=1}^{k} \frac{1}{\gamma N}z_j^2 \\
&= \sqrt{\frac{2}{\gamma}}\sqrt{\frac{1}{N}} \sum_{j=1}^{k} z_j - \frac{1}{\gamma}\frac{1}{N}\sum_{j=1}^{k} z_j^2,
\end{aligned}
$$

where the approximation is based on a second order Taylor expansion of the logarithm and makes sense for large values of N. Introduce the processes $W^N = \{W_t^N, t \in [0, 1]\}$ and $A^N = \{A_t^N, t \in [0, 1]\}$ by

$$
W_t^N = \sqrt{\frac{1}{N}} \sum_{j \le Nt} z_j,
$$

$$
A_t^N = \frac{1}{N} \sum_{j \le Nt} z_j^2.
$$

By Donsker's theorem, one has weak convergence of W^n to W, a standard Brownian motion. Furthermore, A^N converges uniformly in probability to A, $A_t = t$. It follows that $\sqrt{\frac{2}{\gamma}}W^N - \frac{1}{\gamma}A^N$ weakly converges to $\sqrt{\frac{2}{\gamma}}W - \frac{1}{\gamma}A$. Note that A is the quadratic variation process $\langle W \rangle$. Hence $\sqrt{\frac{2}{\gamma}}W^N - \frac{1}{\gamma}A^N$ converges weakly to $\sqrt{\frac{2}{\gamma}}W - \frac{1}{2}\langle\sqrt{\frac{2}{\gamma}}W\rangle$.

Consequently, assuming tN is an integer, for $k = tN$ one finds that the distribution of $\log \theta_k - \log \theta_0$ is approximately (for large N) that of $\sqrt{\frac{2}{\gamma}}W_t - \frac{1}{\gamma}t$, which is normal $N(-\frac{t}{\gamma}, \frac{2t}{\gamma}) = N(-\frac{k}{\gamma N}, \frac{2k}{\gamma N})$. The θ_k can also approximately be generated by the recursion, yielding log-normal random variables,

$$
\theta_{k+1} = \theta_k \exp(\xi_k),
$$

where the ξ_k are IID random variables with common $N(-\frac{1}{\gamma N}, \frac{2}{\gamma N})$ distribution. Indeed for this recursion one finds $\mathbb{E}_k\theta_{k+1} = \theta_k$ and $\mathbb{V}_k(\theta_{k+1}) = \theta_k^2(\exp(\frac{2}{\gamma N}) - 1) \approx \frac{2\theta_k^2}{\gamma N}$, which coincide with the earlier found expressions (14) and (15) for conditional mean and variance.

Moreover the continuous time approximation Z of the θ_k/θ_1, with Z as in (13), is the Doléans exponential $\mathscr{E}\left(\sqrt{\frac{2}{\gamma}}W\right)$ and thus satisfies the stochastic differential equation

$$dZ_t = \sqrt{\frac{2}{\gamma}} Z_t \, dW_t. \tag{16}$$

In Sect. 4.2 we will provide results for the Heston model with the product of θ_1 and this limit process as the squared volatility process.

4 Synthetic data examples

In this section we test the practical performance of our method on challenging synthetic data examples. The goal is to illustrate the ability of our method to recover the volatility in a controlled setting where the ground truth is known and thus the quality of inferential results can be assessed directly. We also show good practical performance of the method in a situation which formally does not fall under the derivations made in Sect. 2; see Sect. 4.2 below.

The simulation setup is as follows: we take the time horizon $T = 1$ and generate $n = 4\,000$ observations as follows. First, using a fine grid of $10n + 1$ time points which are sampled from the Uniform$(0, 1)$ distribution, conditional on including 0 and 1, a realisation of the process X is obtained via Euler's scheme, see Glasserman (2004) or Kloeden and Platen (1992). The n time points $\{t_i\}$ are then taken as a random subsample of those times, conditional on including 1. The settings used for the Gibbs sampler are given in Appendix A.4. In each example below, we plot the posterior mean and 95% marginal credible band (obtained from the central posterior intervals for the coefficients $\xi_k = \sqrt{\theta_k}$ in (6)). The former gives an estimate of the volatility, while the latter provides a means for uncertainty quantification.

All the computations are performed in the programming language Julia, see Bezanson et al. (2017), and we provide the code used in our examples. The hardware and software specifications for the MacBook used to perform simulations are: CPU: Intel(R) Core(TM) M-5Y31 CPU @ 0.90GHz; OS: macOS (x86_64-apple-darwin14.5.0).

4.1 Fan and Gijbels function

Suppose the volatility function is given by

$$s(t) = 3/2 + \sin(2(4t - 2)) + 2\exp(-16(4t - 2)^2), \quad t \in [0, 1]. \tag{17}$$

This is a popular benchmark in nonparametric regression, which we call the Fan & Gijbels function (see Fan & Gijbels 1995). This function was already used in the volatility estimation context in Gugushvili et al. (2020). We generated data using the drift coefficient $b(x) = -x$. For the noise level we took $\eta_v = 0.01$, which is substantial. Given the general difficulty of learning the volatility from noisy observations, one cannot expect to infer it on a very fine resolution (cf. the remarks in Sect. 2.2), and thus we opted for $N = 40$ bins. Experimentation showed that the results were rather robust w.r.t. the number of bins.
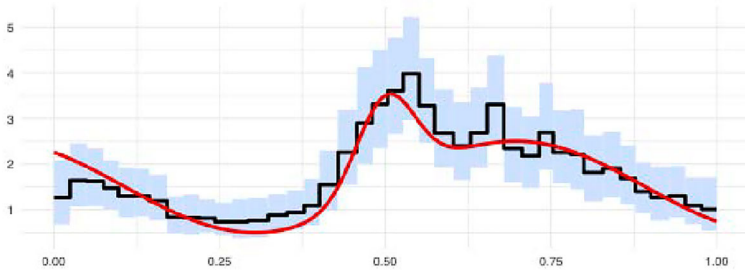
Fig. 1 Posterior and pointwise 95% credible band for the example of Sect. 4.1. The true volatility function is plotted in red, the black step function gives the posterior mean, and the credible band is shaded in blue (color figure online)
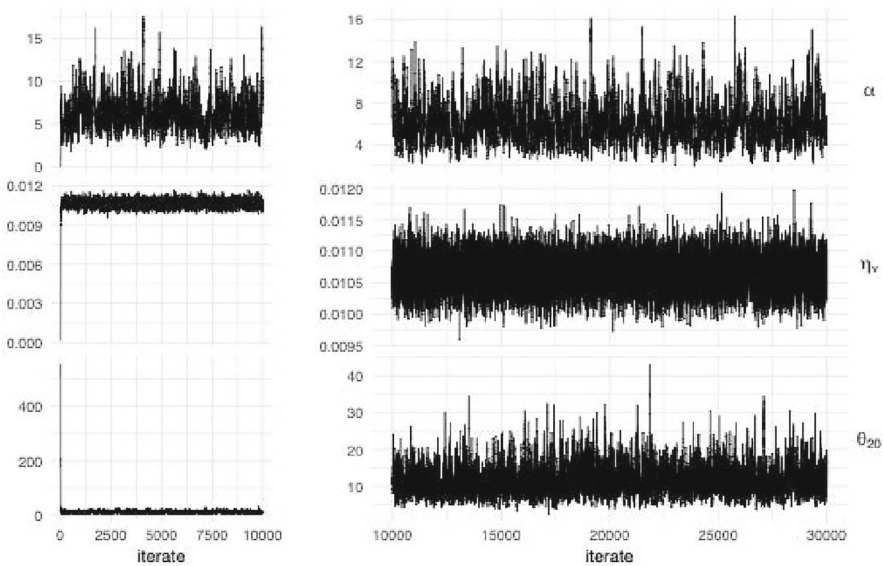

Fig. 2 Trace plots for the Fan and Gijbels function of Sect. 4.1. Left: first 10,000 samples. Right: subsequent 20,000 samples. Top: α, middle: η_v, bottom: θ_{20}

Inference results are reported in Fig. 1. It can be seen from the plot that we succeed in learning the volatility function. Note that the credible band does not cover the entire volatility function, but this had to be expected given that this is a marginal band. Quick mixing of the Markov chains can be visually assessed via the trace plots in Fig. 2. The algorithm took about 11 min to complete.

4.2 Heston model

The Heston model (see Heston 1993, or (Filipovic 2009, Section 10.3.3) and (Brigo & Mercurio 2006, Chapter 19, Appendix A)) is a widely used stochastic volatility model. It stipulates that the price process of a certain asset, denoted by S, evolves over time according to the SDE

$$dS_t = \mu S_t dt + \sqrt{Z_t} S_t dW_t,$$

where the process Z follows the CIR or square root process (see Cox et al. 1985),

$$dZ_t = \kappa(\theta - Z_t)dt + \sigma\sqrt{Z_t}dB_t. \tag{18}$$

Here W and B are correlated Wiener processes with correlation ρ. Following a usual approach in quantitative finance, we work not with S directly, but with its logarithm $X_t = \log S_t$. According to Itô's formula it obeys a diffusion equation with volatility $s(t) = \sqrt{Z_t}$,

$$dX_t = \left(\mu - \frac{1}{2}Z_t\right)dt + \sqrt{Z_t}dW_t.$$

We assume high-frequency observations on the log-price process X with additive noise $V_i \sim N(0, \eta_v)$. In this setup the volatility s is a random function. We assume no further knowledge of the data generation mechanism. This setting is extremely challenging and puts our method to a serious test. To generate data, we used the parameter values $\mu = 0.05$, $\kappa = 7$, $\theta = 0.04$, $\sigma = 0.6$, $\rho = -0.6$, that mimic the ones obtained via fitting the Heston model to real data (see Table 5.2 in van der Ploeg 2006). Furthermore, the noise variance was taken equal to $\eta_v = 10^{-6}$. The latter choice ensures a sufficiently large signal-to-noise ratio in the model (5), that can be quantified via the ratio w_i/η_v of the *intrinsic* noise level w_i to the *external* noise level η_v. Finally, the parameter choice $\mu = 0.04$ corresponds to a 4% log-return rate, which is a reasonable value.

We report inference results with $N = 80$ bins in Fig. 3. These are surprisingly accurate, given a general difficulty of the problem and the amount of assumptions that went into the learning procedure. Note that the number of bins to get accurate results is higher than in the previous example. This is due to the fact that the volatility here is much rougher, Hölder continuous of order less than $\frac{1}{2}$. The Markov chains mix rapidly, as evidenced by the trace plots in Fig. 4. The simulation run took about 12 min to complete. Finally, we note that the Heston model does not formally fall into the framework under which our Bayesian method was derived (deterministic volatility function s). We expect that a theoretical validation why our approach also works in this case requires the use of intricate technical arguments, which lie beyond the scope of the present, practically-oriented paper.

We have also run some experiments for a variation of the above Heston model with the original CIR squared volatility according to (18) replaced with a squared volatility equal to the limit process $\theta_1 Z$, where Z is as in (13), equivalently as in (16), with $\sqrt{\frac{2}{\gamma}} = 0.6$. The experiments showed the influence of the number of bins N and the starting values, illustrated by Figs. 5 and 6. One sees that the results with the lower number of bins $N = 40$ combined with a low initial value are less satisfactory than with the higher $N = 80$ and higher starting value.

Fig. 3 Posterior mean and pointwise 95% credible band for the example of Sect. 4.2. The true volatility function is plotted in red, the black step function gives the posterior mean, and the credible band is shaded in blue (color figure online)
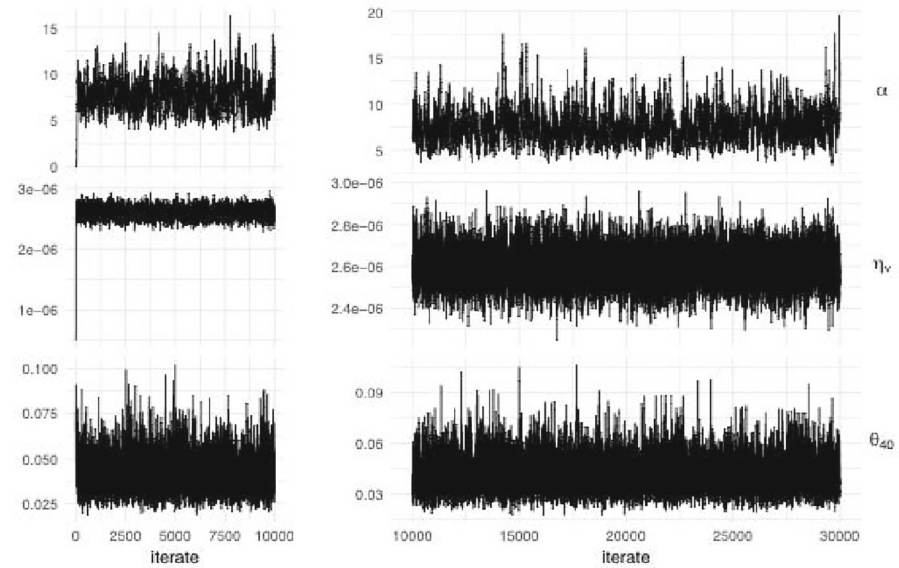

Fig. 4 Trace plots for the Heston model of Sect. 4.2. Left: first 10,000 samples. Right: subsequent 20,000 samples. Top: α, middle: η_v, bottom: θ_{40}
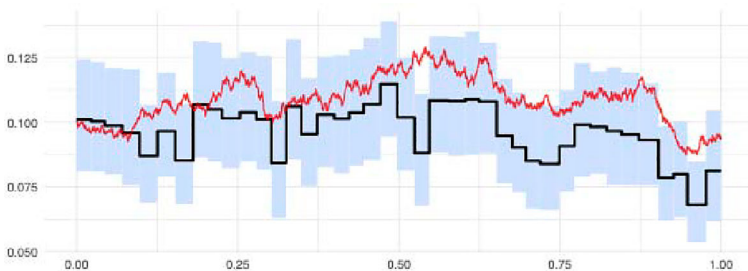

Fig. 5 Posterior results for the Heston model where the CIR volatility is replaced by the root of the continuous time limit of the prior, $N = 40$, starting value of the volatility is 0.1. The colors have the same meaning as in Fig. 3 (color figure online)
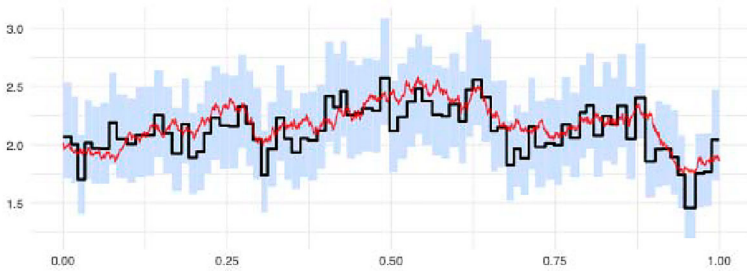
Fig. 6 Posterior results for the Heston model where the CIR volatility is replaced by the root of the continuous time limit of the prior, $N = 80$, starting value of the volatility is 2.0

5 Exchange rate data

Unlike daily stock quotes or exchange rate series that can easily be obtained via several online resources (e.g., Yahoo Finance), high frequency financial data are rarely accessible for free for academia. In this section, we apply our methodology to infer volatility of the high frequency foreign exchange rate data made available by Pepperstone Limited, the London based forex broker.[1] As we shall see below, the inferred volatility looks plausible, and while there is substantial uncertainty surrounding the inferential results left (as quantified by the marginal credible band), nontrivial conclusions can nevertheless be drawn.

Specifically, we use the EUR/USD tick data (bid prices) for 2 March 2015. As the independent additive measurement error model (2) becomes harder to justify for highly densely spaced data, we work with the subsampled data, retaining every 10th observation. Applying various amounts of subsampling is a common strategy in this context; see, e.g., Section 5 in Mancini et al. (2015) for a practical example. In our case subsampling results in a total of $n = 13\,025$ observations over one day, about 9 per minute. See (Mykland & Zhang 2012, Section 2.5) for further motivation and explanation of subsampling and (Zhang et al. 2005, Section 1.2) where it is shown that subsampling (and averaging) is motivated by a substantial decrease in the bias of their estimator. As in Sect. 4.2, we apply the log-transformation on the observed time series, and assume the additive measurement error model (2). The data are displayed in Fig. 7, top panel.

Financial time series often contain jumps, accurate detection of which is a delicate task. As this section serves illustrative purposes only, for simplicity we ignore jumps in our analysis. For high frequency data, volatility estimates are expected to recover quickly after infrequent jumps in the underlying process. This should in particular be the case for our learning procedure, given that we model the volatility as a piecewise constant function, which can be viewed as an estimate of the "histogramised" true volatility. Indeed, effects of infrequent jumps in the underlying process are likely to

---

[1] As of 2020, data are not available from the Pepperstone website any more, but can be obtained directly from the present authors. The data are stored as csv files, that contain the dates and times of transactions and bid and ask prices. The data over 2019 are available for download (after a free registration) at https://www.truefx.com/truefx-historical-downloads.
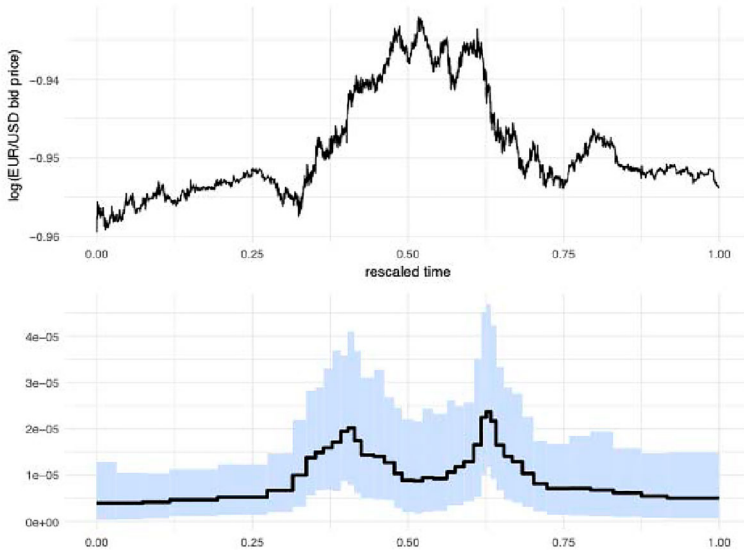
Fig. 7 Top: Natural logarithm of the EUR/USD exchange rate data for 2 March 2015 analysed in Sect. 5. Bottom: Posterior mean (black curve) and pointwise 95% credible band (blue band) for the volatility function. The time axis is rescaled to [0, 1] (color figure online)

get smoothed out by averaging (we make no claim of robustness of our procedure with respect to possible presence of large jumps in the process X, or densely clustered small jumps). An alternative here would have been to use a heuristic jump removal technique, such as the ones discussed in Sabel et al. (2015); next one could have applied our Bayesian procedure on the resulting "cleaned" data.

We report inference results in the bottom panel of Fig. 7, where time is rescaled to the interval [0, 1], so that $t = 0$ corresponds to the midnight and $t = 0.5$ to noon. We used $N = 96$ bins. As seen from the plot, there is considerable uncertainty in volatility estimates. Understandably, the volatility is lower during the night hours. It also displays two peaks corresponding to the early morning and late afternoon parts of the day. Finally, we give several trace plots of the generated Markov chains in Fig. 9. The algorithm took about 33 min to complete. In Figure 8 we give inference results obtained via further subsampling of the data, retaining 50% of the observations. The posterior mean is quite similar to that in Fig. 7, whereas the wider credible band reflects greater inferential uncertainty due to a smaller sample size. The figure provides a partial validation of the model we use.

6 Discussion

In this paper we studied a practical nonparametric Bayesian approach to volatility learning under microstructure noise. From the statistical theory point of view, the problem is much more difficult than volatility learning from noiseless observations. Hence,
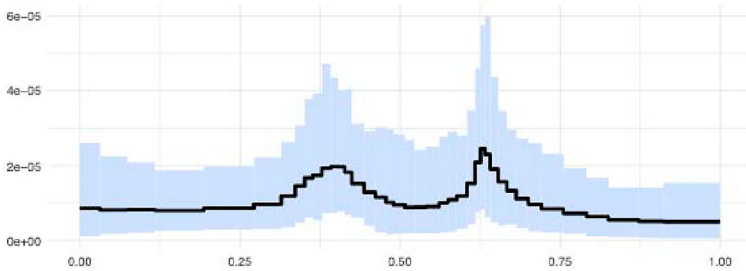
Fig. 8 Posterior mean (black curve) and pointwise 95% credible band (blue band) for the volatility of the further subsampled EUR/USD exchange rate data analysed in Sect. 5. The time axis is rescaled to [0, 1] (color figure online)
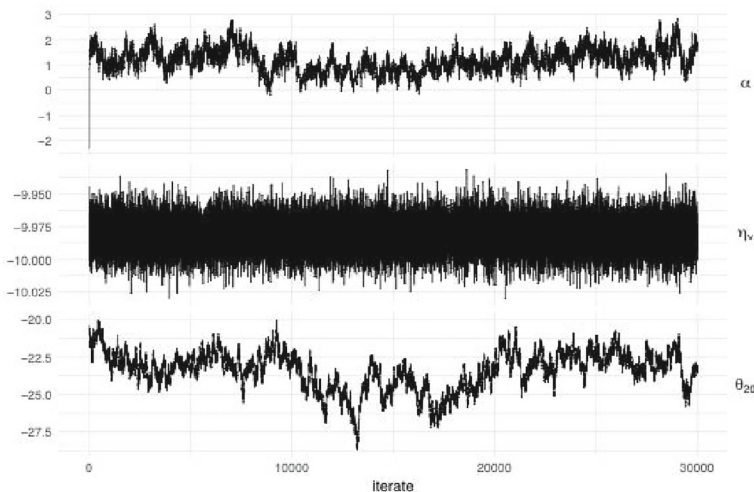

Fig. 9 Log trace plots for the exchange rate data example of Sect. 5. Top: α, middle: η_v, bottom: θ_{20}

accurate inference on volatility under microstructure noise requires large amounts of data. Fortunately, these are available in financial applications. On the other hand, design of a learning method that scales well with data becomes important. Our specification of the prior and a deliberate, but asymptotically harmless misspecification of the drift by taking $b \equiv 0$ are clever choices that enable us to combine our earlier work in Gugushvili et al. (2019b) with the FFBS algorithm for Gaussian linear state space models. This gives a conceptually simple and fast algorithm (Gibbs sampler) to obtain samples from the posterior, from which inferences can be drawn in a straightforward way. A very interesting topic for future research within our approach is to explicitly account for the possible presence of jumps in financial time series.

Code availability The computer code to reproduce numerical examples in this article is available at Gugushvili et al. (2022).

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

A Details on update steps in the Gibbs sampler

A.1 Drawing $x_{0:n}$

We first describe how to draw the state vector $x_{0:n}$ conditional on all other parameters in the model and the data $y_{1:n}$. Note that for u_i in (5) we have by (4) that $u_i \sim N(0, w_i)$, where

$$w_i = \theta_k \Delta_i, \quad i \in [(k-1)m+1, km], \quad k = 1, \dots, N-1,$$
$$w_i = \theta_N \Delta_i, \quad i \in [(N-1)m+1, n]. \tag{19}$$

By Eq. (4.21) in Petris et al. (2009) (we omit dependence on $\theta_{1:N}$, η_v in our notation, as they stay fixed in this step),

$$p(x_{0:n}|y_{1:n}) = \prod_{i=0}^{n} p(x_i|x_{i+1:n}, y_{1:n}),$$

where the factor with $i = n$ in the product on the righthand side is the filtering density $p(x_n|y_{1:n})$. This distribution is in fact $N(\mu_n, C_n)$, with the mean μ_n and variance C_n obtained from Kalman recursions

$$\mu_i = \mu_{i-1} + K_i e_i, \quad C_i = K_i \eta_v, \quad i = 1, \dots, n.$$

Here

$$K_i = \frac{C_{i-1} + w_i}{C_{i-1} + w_i + \eta_v}, \quad i = 1, \dots, n,$$

is the *Kalman gain*. Furthermore, $e_i = y_i - \mu_{i-1}$ is the one-step ahead *prediction error*, also referred to as *innovation*. See Petris et al. (2009), Section 2.7.2. This constitutes the forward pass of the FFBS.

Next, in the backward pass, one draws backwards in time $\widetilde{x}_n \sim N(\mu_n, C_n)$ and $\widetilde{x}_{n-1}, \dots \widetilde{x}_0$ from the densities $p(x_i|\widetilde{x}_{i+1}, y_{1:n})$ for $i = n-1, n-2, \dots, 0$. It holds that $p(x_i|\widetilde{x}_{i+1:n}, y_{1:n}) = p(x_i|\widetilde{x}_{i+1}, y_{1:n})$, and the latter distribution is $N(h_i, H_i)$, with

$$h_i = \mu_i + \frac{C_i}{C_i + w_{i+1}}(\widetilde{x}_{i+1} - \mu_i), \quad H_i = \frac{C_i w_{i+1}}{C_i + w_{i+1}}.$$

For every i, these expressions depend on a previously generated $\widetilde{x}_{i+1}$ and other known quantities only. The sequence $\widetilde{x}_0, \widetilde{x}_1, \ldots, \widetilde{x}_n$ is a sample from $p(x_{0:n}|y_{1:n})$. See Section 4.4.1 in Petris et al. (2009) for details on FFBS.

A.2 Drawing η_v, $\theta_{1:N}$ and $\zeta_{2:N}$

Using the likelihood expression from Sect. 2.3 and the fact that $\eta_v \sim \text{IG}(\alpha_v, \beta_v)$, one sees that the full conditional distribution of η_v is given by

$$\eta_v|x_{1:n}, y_{1:n} \sim \text{IG}\left(\alpha_v + \frac{n}{2}, \beta_v + \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i)^2\right).$$

Similarly, using the likelihood expression from Sect. 2.3 and the conditional distributions in (7), one sees that the full conditional distributions for $\theta_{1:N}$ are

$$\theta_1|\zeta_2, x_{1:n} \sim \text{IG}\left(\alpha_1 + \alpha + \frac{m_1}{2}, \beta_1 + \frac{\alpha}{\zeta_2} + \frac{Z_1}{2}\right),$$

$$\theta_k|\zeta_k, \zeta_{k+1}, x_{1:n} \sim \text{IG}\left(2\alpha + \frac{m_k}{2}, \frac{\alpha}{\zeta_k} + \frac{\alpha}{\zeta_{k+1}} + \frac{Z_k}{2}\right), \quad k = 2, \ldots, N-1,$$

$$\theta_N|\zeta_N, x_{1:n} \sim \text{IG}\left(\alpha + \frac{m_N}{2}, \frac{\alpha}{\zeta_N} + \frac{Z_N}{2}\right).$$

The full conditional distributions for $\zeta_{2:N}$ are

$$\zeta_k|\theta_k, \theta_{k-1} \sim \text{IG}\left(2\alpha, \frac{\alpha}{\theta_{k-1}} + \frac{\alpha}{\theta_k}\right), \quad k = 2, \ldots, N.$$

A.3 Drawing α

The unnormalised full conditional density of α is

$$q(\alpha) = \pi(\alpha)\left(\frac{\alpha^\alpha}{\Gamma(\alpha)}\right)^{2(N-1)} \exp\left(-\alpha\sum_{k=2}^{N}\frac{1}{\zeta_k}\left(\frac{1}{\theta_{k-1}} + \frac{1}{\theta_k}\right)\right)\prod_{k=2}^{N}\left(\theta_{k-1}\theta_k\zeta_k^2\right)^{-\alpha}.$$

The corresponding normalised density is nonstandard, and the Metropolis-within-Gibbs step (see, e.g., Tierney 1994) is used to update α. The specific details are exactly the same as in Gugushvili et al. (2019b).

A.4 Gibbs sampler

Settings for the Gibbs sampler in Sect. 4 are as follows: we used a vague specification $\alpha_1, \beta_1 \to 0$, and also assumed that $\log \alpha \sim N(1, 0.25)$ and $\eta_v \sim \text{IG}(0.3, 0.3)$ in Sect. 4.1. For the Heston model in Sect. 4.2 we used the specification $\eta_v \sim$

IG(0.001, 0.001). Furthermore, we set $x_0 \sim N(0, 25)$. The Metropolis-within-Gibbs step to update the hyperparameter α was performed via an independent Gaussian random walk proposal (with a correction as in Wilkinson (2012)) with scaling to ensure the acceptance rate of about $30 - 50\%$. The Gibbs sampler was run for $30\,000$ iterations, with the first third of the samples dropped as burn-in.

References

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59*(1), 65–98.

Brigo, D., & Mercurio, F. (2006). *Interest rate models-theory and practice: With smile, inflation and credit* (Vol. 2). Springer.

Cemgil, A. T., & Dikmen, O. (2007). Conjugate gamma Markov random fields for modelling nonstationary sources. In M. E. Davies, C. J. James, S. A. Abdallah, & M. D. Plumbley (Eds.), *Independent component analysis and signal separation: 7th International Conference, ICA 2007, London, UK, September 9–12, 2007. Proceedings* (pp. 697–705). Springer.

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica, 53*(2), 385–407.

Fan, J., & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of Royal Statistics Society Series B, 57*(2), 371–394.

Filipovic, D. (2009). *Term-structure models. A graduate course*. Springer.

Ghosal, S., & van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference, Cambridge series in statistical and probabilistic mathematics* (Vol. 44). Cambridge: Cambridge University Press.

Glasserman, P. (2004). *Monte Carlo methods in financial engineering* (Vol. 53). Springer.

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2019). Fast and scalable non-parametric Bayesian inference for Poisson point processes. RESEARCHERS.ONE. https://www.researchers.one/article/2019-06-6

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2019). Nonparametric Bayesian volatility estimation. In J. de Gier, C. E. Praeger, & T. Tao (Eds.), *2017 MATRIX annals* (pp. 279–302). Springer.

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2020). Nonparametric Bayesian estimation of a Hölder continuous diffusion coefficient. *Brazilian Journal of Probability and Statistics, 34*(3), 537–579. https://doi.org/10.1214/19-BJPS433.

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2022). *Julia code for nonparametric Bayesian volatility learning under microstructure noise*. https://doi.org/10.5281/zenodo.6801410

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies, 6*(2), 327–343.

Hoffmann, M., Munk, A., & Schmidt-Hieber, J. (2012). Adaptive wavelet estimation of the diffusion coefficient under additive error measurements. *Annales de l'IHP Probabilités et Statistiques, 48*(4), 1186–1216.

Ignatieva, K., & Platen, E. (2012). Estimating the diffusion coefficient function for a diversified world stock index. *Computer Statistics Data Analysis, 56*(6), 1333–1349.

Jacod, J., Li, Y., Mykland, P., Podolskij, M., & Vetter, M. (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes Application, 119*(7), 2249–2276.

Jacod, J., & Shiryaev, A. (2013). *Limit theorems for stochastic processes* (Vol. 288). Springer.

Kanaya, S., & Kristensen, D. (2016). Estimation of stochastic volatility models by nonparametric filtering. *Econometric Theory, 32*(4), 861–916.

Kloeden, P. E., & Platen, E. (1992). *Numerical solution of stochastic differential equations, Applications of Mathematics (New York)* (Vol. 23). Springer. https://doi.org/10.1007/978-3-662-12616-5.

Mancini, C., Mattiussi, V., & Renò, R. (2015). Spot volatility estimation using delta sequences. *Finance and Stochastics, 19*(2), 261–293.

Müller, P., & Mitra, R. (2013). Bayesian nonparametric inference—Why and how. *Bayesian Analyses, 8*(2), 269–302.

Müller, P., Quintana, F. A., Jara, A., & Hanson, T. (2015). *Bayesian nonparametric data analysis. Springer Series in Statistics*. Springer.

Munk, A., & Schmidt-Hieber, J. (2010). Lower bounds for volatility estimation in microstructure noise models. *Borrowing strength: Theory powering applications—A Festschrift for Lawrence D. Brown, Inst. Math. Stat. (IMS) Collect.* (Vol. 6, pp. 43–55). Institute of Mathematics Statistics.

Musiela, M., & Rutkowski, M. (2005). Martingale methods in financial modelling. *Stochastic modelling and applied probability* (2nd ed., Vol. 36). Springer.

Mykland, P. A., & Zhang, L. (2009). Inference for continuous semimartingales observed at high frequency. *Econometrica, 77*(5), 1403–1445.

Mykland, P. A., & Zhang, L. (2012). The econometrics of high-frequency data. In C. R. C. Press (Ed.), *Statistical methods for stochastic differential equations, Monogr. Statist. Appl. Probab.* (Vol. 124, pp. 109–190). CRC Press.

Papaspiliopoulos, O., Roberts, G. O., & Stramer, O. (2013). Data augmentation for diffusions. *Journal of Computational and Graphical Statistics, 22*(3), 665–688.

Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic linear models with R. Use R!* Springer.

Reiß, M. (2011). Asymptotic equivalence for inference on the volatility from noisy observations. *Annals of Statistics, 39*(2), 772–802.

Sabel, T., Schmidt-Hieber, J., & Munk, A. (2015). Spot volatility estimation for high-frequency data: Adaptive estimation in practice. *Modeling and stochastic learning for forecasting in high dimensions, Lect. Notes Stat.* (Vol. 217, pp. 213–241). Springer.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis. Monographs on statistics and applied probability*. Chapman & Hall.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of American Statistics Association, 82*(398), 528–550 (**With discussion and with a reply by the authors**).

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics, 22*(4), 1701–1762.

van der Ploeg, A. P. C. (2006). Stochastic volatility and the pricing of financial derivatives. Ph.D. thesis, University of Amsterdam.

van der Meulen, F., & Schauer, M. (2017). Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronics Journal of Statistics, 11*(1), 2358–2396.

Wilkinson, D. J. (2012). Metropolis Hastings MCMC when the proposal and target have differing support. https://darrenjw.wordpress.com/2012/06/04/metropolis-hastings-mcmc-when-the-proposal-and-target-have-differing-support/. Accessed 23 Dec 2017

Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association, 100*(472), 1394–1411.