Okhrin, Ostap; Rockinger, Michael; Schmid, Manuel

Article — Published Version

Observations concerning the estimation of Heston's stochastic volatility model using HF data

Statistical Papers

Suggested Citation: Okhrin, Ostap; Rockinger, Michael; Schmid, Manuel (2025) : Observations concerning the estimation of Heston's stochastic volatility model using HF data, Statistical Papers, ISSN 1613-9798, Springer Berlin Heidelberg, Berlin/Heidelberg, Vol. 66, Iss. 4, https://doi.org/10.1007/s00362-025-01710-0

This Version is available at:
https://hdl.handle.net/10419/323270

Statistical Papers (2025) 66:90
https://doi.org/10.1007/s00362-025-01710-0

REGULAR ARTICLE

Observations concerning the estimation of Heston's stochastic volatility model using HF data

Ostap Okhrin[1] · **Michael Rockinger**[2] · **Manuel Schmid**[3]

Received: 29 October 2024 / Revised: 13 January 2025 / Published online: 8 May 2025

Abstract

This paper presents a comprehensive simulation study on estimating parameters for the popular Heston stochastic volatility model. Leveraging high-frequency data, we explore, in a data-science type exercise, various spot-volatility estimation and sampling techniques, improving existing methods to enhance parameter accuracy. Through extensive simulations, we report difficulties in generating correct parameter estimates for realistic parameter settings where the volatility dynamic does not satisfy the Feller condition. This study contributes valuable insights into the practical implementation of the Heston model and its applicability to high-frequency data. We find that the scheme of Azencott et al. (2020) with uniform kernel weighting provides reliable and efficient parameter estimates. It is advised to also apply a Jackknife estimation to corroborate the findings.

Keywords Spot volatility · Square-root process · Method of moments · Maximum-likelihood · High-frequency data

The views expressed in this paper are those of the authors and are not necessarily reflective of views from their employers. All simulations and estimations were performed with the open-source software R. We are most grateful to the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) who partially funded the project under number 497262150.

✉ Ostap Okhrin
ostap.okhrin@tu-dresden.de

Michael Rockinger
michael.rockinger@unil.ch

Manuel Schmid
manuel.m.schmid@deutschebahn.com

1 Technische Universität Dresden, Institute of Economics and Transport, School of Transportation, Chair of Statistics and Econometrics esp. Transportation, Dresden, Germany

2 University of Lausanne, Faculty of Business and Economics, Lausanne, CH 1015, Switzerland

3 Deutsche Bahn AG, Poststraße 20, 60329 Frankfurt am Main, Germany

1 Introduction

The paper presents an extensive simulation study focused on estimating parameters in the Heston stochastic volatility model, Heston (1993) and of the volatility process modeled as a Cox-Ingersoll-Ross process, (Cox, Ingersoll and Ross, 1985). We simulate data corresponding to high-frequency observations and attempt to accurately estimate the parameters of this model, especially in practical scenarios where volatility does not satisfy the Feller condition. In this realistic case, volatility can approach zero in finite time, complicating parameter estimation and increasing the risk of bias in methods reliant on conditional moments.

The idea of using high-frequency data to estimate the spot volatility goes back to work by Andersen et al. (2001); Barndorff-Nielsen and Shephard (2002), who developed estimates using sums of squared returns, also named realized volatility. The work by Andersen and Lund (1997); Chib et al. (2002) was pathbreaking because they demonstrated that volatility models could be estimated on raw returns. Andersen et al. (2003); Bollerslev and Zhou (2002) went one step further by using integrated volatility as a proxy for spot volatility to estimate the dynamic of volatility. The idea of filtering realized volatility using kernels appears to go back to Renò (2006), then to Kanaya and Kristensen (2016).

A couple of years later, the work by Bandi and Renò (2016, 2018) suggested taking spot volatility and sampling it at relatively distant periods. For instance, one uses observations every 20 s to estimate volatility over a window of one hour; this yields one estimate of spot volatility. One repeats this process one day later. The daily sampled volatility estimates will then be used to estimate some parametric or non-parametric models. Observations sampled in this way are called pseudo-observations. This research leaves, however, open the question of what an optimal period for the pseudo sampling is. Research by Azencott et al. (2020) (ART) indicates how to select the frequency of returns for the spot volatility estimation and at which frequency this spot volatility estimate should be sub-sampled for the estimation of a parametric model. They also provide closed-form solutions for the parameter estimates based on the method of moments for Heston's volatility-process parameters. Then no iterative method like in Generalized Method of Moments (GMM) or Maximum Likelihood (ML) estimation is required.

In this paper, we want to summarize an extensive empirical exercise examining the quality of various spot-volatility estimates in a simulation setting. Additional results on the simulation and an application to actual data is provided in an online appendix. We also want to investigate if it is possible to estimate correctly the parameters of Heston's model, i.e. the stochastic volatility dynamic and the correlation between innovations. Our first contribution combines ART's pseudo sampling scheme with spot volatility estimation, Kanaya and Kristensen (2016); Bandi and Renò (2018). We study the quality of various weighting schemes in this setting and find that a uniform scheme outperforms the Gaussian, Epanechnikov, and a one-sided Kernel. We also find that the pseudo-sampling scheme of ART dominates other ad-hoc schemes.

Next, we perform an extensive simulation exercise involving 18 realistic parameter vectors for the volatility dynamic. We use the ART method of moments as well as 12 GMM and Minimum Distance estimation methods. We also consider 6 ML or Pseudo

ML approaches. The key finding is that the estimation scheme of Shoji and Ozaki (1998), possibly with a jackknive extension, is particularly successful. We also find that the method of ART is fast and yields stable parameter estimations.

As we inspect the quality of the parameter estimates of the stochastic volatility dynamic, we find that the long-run variance parameter is most straightforward to estimate, which is unsurprising since this parameter does not require a model and is essentially estimated as a mean of squared log-returns. We find that parameter constellations where the process may come near 0 will yield worse estimates of the mean reversion and the volatility of the volatility parameter. Conditions for the variance process to come close to zero are a low long-term variance, a high variance variability, and a more persistent mean reversion. A low Feller ratio will be indicative of such conditions.

An additional approach is the Markov Chain Monte Carlo (MCMC) estimation method. Among others (Eraker et al. 2003; Jacquier et al. 2004; Johannes and Polson 2010) discuss theoretical aspects. We also implemented this methodology in a preliminary exercise. Unfortunately, we found this method computationally costly (we perform thousands of estimations in our simulations with long time series) and decided to restrict ourselves to non-Bayesian estimation.

In summary, this paper contributes to the literature on how to estimate the parameter's of Heston's model, a notoriously difficult exercise. In it we: (i) combine Azencott et al. (2020) scheme and bucket-based approaches for spot variance estimation, (ii) rigorously evaluate a list of estimation methods under realistic parameter settings, and (iii) provide recommendations for parameter estimation in HF financial modeling. Those estimates are interesting per se to understand the underlying asset dynamic. They can help improve Value at Risk and Expected shortfall estimates for porfolios containing a stock. They can be used to forecast volatility wich is relevant for hedging strategies or optimal trading strategies. Last, for situations with scarce or noisy option data, our investigation provides an alternative approach to obtain parameters.

In the next section, we describe the data-generating process. In Sect. 3 we indicate how to estimate the spot variance. In Sect. 4, we indicate the methods we use to estimate the parameters of the model. In Sect. 5 we discuss the results. Finally, 6 concludes this paper.

2 Main data generating process

Throughout this paper, we will assume that observed (log-)returns are generated by the following system of stochastic differential equations (SDEs):

$$\mathrm{d}S_t = \mu S_t \mathrm{d}t + \sqrt{v_t}\, S_t\, \mathrm{d}W_t^x, \tag{1}$$

$$\mathrm{d}X_t = \left(\mu - \frac{1}{2} v_t\right) \mathrm{d}t + \sqrt{v_t}\, \mathrm{d}W_t^x, \tag{2}$$

$$\mathrm{d}v_t = \kappa\,(\theta - v_t)\,\mathrm{d}t + \sigma\,\sqrt{v_t}\,\mathrm{d}W_t^v, \tag{3}$$

$$\mathbb{E}[\mathrm{d}W_t^x\,\mathrm{d}W_t^v] = \rho\,\mathrm{d}t, \tag{4}$$

where (1) is the price (S_t), (2) is the log-price (X_t) process useful for the construction of log-returns, with $X_t = \log(S_t)$. The terms $\mathrm{d}W_t^x$ and $\mathrm{d}W_t^v$ are increments of Brownian motions that are correlated via (4) with constant correlation $\rho \in (-1, +1)$. The parameter μ controls the drift of the price process. The complete system is known as Heston's model, see Heston (1993). A possible extension with jumps can be found in Bates (1996); Duffie et al. (2000). In this study, we focus on estimating the parameters in Heston's model using high-frequency data. For an estimation on real data, one may have to filter out possible jumps in returns. For this reason, we do not present the extended versions of this model incorporating jumps.

The evolution of latent spot variance v_t is modeled by the well-known Cox Ingersoll Ross (CIR) process (3) [see Cox et al. (1985)], where κ is the speed of mean reversion, θ is the long-run mean, and σ is the volatility of variance. In addition, as shown already by Feller (1951), if the Feller ratio defined by $2\kappa\theta/\sigma^2 \geq 1$, which is the Feller condition, then the process will not reach 0. For $0 \leq 2\kappa\theta/\sigma^2 < 1$, the origin will be reached in finite time but is instantaneously reflecting, and for the theoretical case where $\theta = 0$, the origin is absorbing.

From Cox et al. (1985), the expression for the first two conditional moments is also known, $\mathbb{E}[v_t|v_0] = v_0\, e^{-\kappa t} + \theta\left(1 - e^{-\kappa t}\right)$, and $\mathbb{V}[v_t|v_0] = \sigma^2\theta\left(1 - e^{-\kappa t}\right)^2/2\kappa + v_0\sigma^2\left(e^{-\kappa t} - e^{-2\kappa t}\right)/\kappa$. The unconditional mean is θ, and the unconditional variance is $\sigma^2\theta/2\kappa$, as can be seen by letting $t \to \infty$. A historical discussion and expressions for conditional and unconditional higher(co-) moments to this process can be found in Okhrin et al. (2022).

3 Spot variance estimation

We will assume the availability of n equidistant, discrete sampled observations of log-prices generated by the system of stochastic differential equations (2) - (4) over the time span $[0, T]$. Therefore, the time between two observations is $\Delta_{HF} = T/n$ for a large n. We will refer to parameter estimation with data sampled with a time span between observations of Δ_{HF} as a high-frequency estimation. In practice this can be seconds up to a few minutes between trades.

Since the spot variance can not be observed, it has to be filtered from the available high-frequency (log)-return data. In the literature, there are mainly two ways to do so. In the first approach, by using kernel-based estimates, one can generate a high-frequency series of spot volatility estimates. In the second approach, one estimates over lower frequency spaced windows points of spot volatility. The volatility process is then estimated with traditional methods using those volatility estimates.

3.1 Continuous spot variance estimation

This approach can be found in Kanaya and Kristensen (2016) who presents a kernel-based approach. With our notations, the spot variance estimator is defined as

$$\hat{v}_j^{KK} = \frac{1}{h}\sum_{k=1}^{n} \mathcal{K}\left[\frac{(k-1)\Delta_{HF} - j\Delta_{HF}}{h}\right] r_k^2, \quad j \in \{1, \cdots, n-1\},$$

where $\mathcal{K}(\cdot)$ is some kernel function, h some bandwidth and $r_k = X_k - X_{k-1}$ the high frequency log-return at step k. Intuitively, in this approach, one uses all available returns to estimate an instantaneous volatility based on adjacent returns. Depending on the kernel, those returns may be rather distant. It results that the spot volatility can be determined at the given (log-)return frequency and, therefore, without any loss in sample size.

There are potentially two issues with this approach. The first one is that since potentially all datapoints have to be considered for evaluating one spot volatility, the numerical effort may become overwhelming. By using truncated windows, one can mitigate this issue. Another issue is that the method requires selection of the bandwidth parameter h, which can be done with cross-validation. The time required to do so in a simulation with thousands of series may represent a significant burden.

3.2 Bucket based spot variance estimation

An alternative approach can be found in Bandi and Renò (2018) who present an approach based on buckets collecting J high-frequency observations and estimating point estimates of variance. In practice, this could involve sampling data every day at a low frequency, such as for a single hour. Within this one-hour window, high-frequency data–for instance, minute-by-minute observations–would then be used to compute the realized volatility.

To illustrate their procedure, let $\Delta_{LF} = m\Delta_{HF}$, $m < n$ define some lower frequency within the sample. That is for each m-th high frequent observation we select a low frequency estimate. Consequently, we will have a sample of $M = T/(m\Delta_{HF}) = n/m$ low-frequency observations. In this approach, the spot variance estimator is constructed using squared returns, as in

$$\hat{v}_i^{BR} = \frac{1}{J\Delta_{HF}}\sum_{k=1}^{J} r_{i,k}^2, \quad i = \{1, \cdots, M\},$$

where J denotes the number of high-frequency data points within the bucket at each low-frequency time step which are effectively used to estimate the variance, and $r_{i,k} = X_{im-J+k} - X_{im-J+k-1}$ is the k-th high-frequency log-return observation within the i-th bucket. It is important to note that Bandi and Renò (2018) recommend choosing $J << m$ and therefore the length of each bucket expressed in physical time is $\xi =$

$J\Delta_{HF} << \Delta_{LF}$. Explicitly, they suggest to only use around $1/7$ of the m high-frequency observations included within a low frequent step. For this reason, define $J_{gap} = m - J > 0$, that is the number of high-frequency observations in each low frequency interval that are not used in the estimation of spot volatility.

Azencott et al. (2020) (ART) starts in very different manner. In their paper, they assume that the econometrician imposes the size of the window over which realized volatility is to be estimated, denoted as above ξ, say one day. They then provide theoretical results on the optimal number of high-frequency observations needed to compute an estimate of realized volatility. Additionally, they derive results on the optimal spacing between realized volatility estimates and the number of such windows required to ensure parameter estimates converge at a known rate. Thus, their approach introduces a structured framework for sampling and highlights the associated trade-offs, thereby extending (Bandi and Renò 2018). One trade-off involves balancing the size of the window used to estimate realized volatility against the total number of observations needed. Since their bounds require an unrealistic large sample size, it is simulation exercises like ours that validate this method for practical purposes.

More formally, let ξ denote the window length over which realized volatility should be estimated. They then demonstrate that the optimal number of high-frequency returns is given by $J_{opt} = 1/\xi^2$.[1]

Furthermore, their paper reveals that the number of low frequency estimates should be $m \sim \xi^{-3/2}$, and $\Delta_{LF} \sim \xi^{1/2}$, with J_{opt} as above, there will be convergence in probability at a speed of convergence proportional to $\sqrt{\xi}$.

Eventually, and as already suggested by ART, we wish to bring together the approaches of Bandi and Renò (2018) by weighting the J return observations used with various kernels. The estimator is then defined as

$$\hat{v}_i^{New} = \frac{1}{J\Delta_{HF}} \sum_{k=1}^{J} w_k r_{i,k}^2, \quad i \in \{1, \cdots, M\}, \tag{5}$$

with $w_k = \frac{J\mathcal{K}(k)}{\sum_{j=1}^{J}\mathcal{K}(j)}$ and $J \leq m$, thus $J_{gap} \geq 0$. The ART (Azencott et al. 2020) paper contributes along two dimensions. The most innovative dimension is the derivation of an optimal sampling scheme relating the high frequency time interval and the high frequency number of observations per bucket and the optimal distance between buckets to generate series of realized volatility. The second contribution is to provide explicit method of moment estimates of the parameters of the volatility dynamics. These estimates are to be found below.

4 Estimating the volatility dynamic: θ, κ and σ

If one has true or estimated observations of the variance process (3) there is a vast range of estimation approaches with which the underlying parameters can be estimated. In this section, we discuss some of the most promising approaches, before conducting an

[1] Since $1/\xi^2$ is a very large number, they empirically show that the smaller $1/\xi$ is also satisfactory.

exhaustive Monte Carlo study to investigate the quality of the discussed estimators. Since the parameter θ is the long-term volatility we estimate this parameter in a separate step as the average volatility. This has the advantage that the estimation only requires the estimation of two additional parameters κ and σ.

4.1 Method of moment approach

Azencott et al. (2020) (ART) derive explicit parameter estimators of the process from the first two unconditional central moments and the unconditional auto-covariance function $K(b) = Cov(v_t, v_{t+b})$, where b is a time lag given by a real number. Introducing

$$\hat{K}(b) = \frac{1}{M-B} \sum_{k=1}^{M-B} \hat{v}_k \hat{v}_{k+B} - \hat{m}^2,$$

where M is the size of a sample and where B is the index of the high-frequency time interval approximating the actual time lag. Formally $B = \left\lfloor \frac{b}{\Delta_{LF}} \right\rfloor$. The estimators are given by $\hat{\kappa} = -\frac{1}{b} \log \frac{\hat{K}(b)}{\hat{K}(0)}$, $\hat{\theta} = \hat{m}$, $\hat{\sigma}^2 = \frac{2\hat{K}(0)\hat{\kappa}}{\hat{\theta}}$, with $\hat{m} = M^{-1} \sum_{i=1}^{M} \hat{v}_i$. The parameter b is a hyper-parameter that represents the time lag in physical time, and, thus, has to be chosen a priori. ART suggests a value of $b \approx 0.6$ independent of Δ_{LF}. In any case, it has to be chosen so that the estimator functions can be evaluated which will only be the case for $\hat{K}(b) > 0$. However, this will not hold for large values of B, that is large values of b.

4.2 GMM approaches

Another branch of estimation methods of the CIR process is given by the GMM approach of Hansen (1982). For example, Bollerslev and Zhou (2002) estimate the process via a GMM procedure using the conditional moments of integrated variance. Eventually, since general and closed form expressions for the (un)conditional (co)moments of the processes solution are known from Cox et al. (1985), and more recently (Okhrin et al. 2022), we will also experiment below with GMM moment conditions that we have not encountered so far in the literature.

Let us briefly review the GMM approach. Assume all the parameters of a model are grouped in a p-component vector $\psi = (\kappa, \sigma)^\top$ and that X_i denotes the vector of data. Define $u_i = u(X_i; \psi)$, for $i = \{1, \ldots, M\}$. Then define an $r \times 1$ vector of moment conditions, where $r \geq p$. It is assumed that $\mathbb{E}[u_i] = 0, \ \forall\, i$. Further, introduce the vector of means $g_M(\psi) = \frac{1}{M} \sum_{i=1}^{M} u_i$. It is assumed that asymptotically, the mean vector converges to a theoretical moment vector, $g_M(\psi) \xrightarrow{a.s.} \mathbb{E}[u_i]$. The objective function of the GMM procedure is defined as

$$\hat{\psi} = \arg\min_{\psi} g_M(\psi)^\top W g_M(\psi), \tag{6}$$

where W is some positive definite weighting matrix. Hansen (1982) determines that the optimal choice for W is S^{-1}, where $S = \mathbb{E}\left[u_i u_i^\top\right]$ is the long-run covariance matrix of u. From Newey and West (1987) it is known that S can be estimated via $\hat{S} = \hat{S}_0 + \sum_{j=1}^{\ell} \omega_j \left(\hat{S}_j + \hat{S}_j^\top\right)$, with weights ω_j usually defined as Bartlett weights $\omega_j = 1 - \frac{j}{\ell+1}$ and $\hat{S}_j = \frac{1}{n}\sum_{i=j+1}^{n} u_i u_{i-j}^\top$, $j = 0, 1, \cdots, \ell$. The maximum lag ℓ has to be chosen beforehand, for example, as described in Newey and West (1994).

Since the computation of S requires knowledge of an estimate of ψ the objective function (6) is maximized in two steps. In a first step, referred to as phase 1, one takes $W = I$, an identity matrix. Denote the resulting parameter vector $\hat{\psi}^{(1)}$ with which S can be estimated. In a second step, also called phase 2, it is possible to get the overall estimator of (6).

Phase 2 can be iterated a certain number of times. In our empirical work, we will stop iterating when the L^2 distance of two consecutive parameter vectors $\hat{\psi}^{(s)}$ does not evolve by more than 10^{-3} or when the algorithm has iterated 30 times. It should be noticed that, in theory, already after a first iteration, asymptotic efficiency has been reached.

For the following, it is helpful to introduce

$$
\begin{aligned}
u_i^{powers}(\kappa,\sigma) &= \left[v_{i+1} - \mathbb{E}[v_{i+1}], \cdots, v_{i+1}^{pow} - \mathbb{E}[v_{i+1}^{pow}]\right]^\top, \\
u_i^{co.pow}(\kappa,\sigma) &= \Big[v_{i+1}^1 v_i^1 - \mathbb{E}[v_{i+1} v_i], \cdots, v_{i+1}^1 v_{i+1-lag}^1 - \mathbb{E}[v_{i+1}^1 v_{i+1-lag}^1], \\
&\quad \cdots, v_{i+1}^{cpo} v_i^1 - \mathbb{E}[v_{i+1}^{cpo} v_i^1], \cdots, v_{i+1}^{cpo} v_{i+1-lag}^1 - \mathbb{E}[v_{i+1}^{cpo} v_{i+1-lag}^1], \\
&\quad \cdots, v_{i+1}^{cpo} v_{i+1-lag}^{cpo} - \mathbb{E}[v_{i+1}^{cpo} v_{i+1-lag}^{cpo}]\Big]^\top, \\
u_i^{moms}(\kappa,\sigma) &= \left[u_i(\kappa,\sigma)^{powers\top}, u_i(\kappa,\sigma)^{co.pow\top}\right]^\top.
\end{aligned}
$$

In the first equation, pow denotes the highest power of volatility to be included in the estimation. We will take $pow = 2$ or 3. In the second equation, we denote by cpo the highest co-power. That is the power to which we will raise the volatilities. If the parameter $cpo = 0$ then, by definition, there is no co-power term. In the last equation, we stack the moments. The true moments are the unconditional moments of the CIR process, which we take from Okhrin et al. (2022), and which depend on the parameters of the process κ and σ. An implementation of the conditional and unconditional moments can be found in a GitHub depository.[2]

With those definitions, we can introduce various estimation methods. To simplify, we focus in terms of notation only on phase 2 estimates from GMM. By removing a `W` in the label of the following methods, one obtains the phase 1 estimates.

`gmmW` denotes the estimate of the GMM as outlined above. In method `gmm` we use the first four powers of the data, i.e., $u_i = [v_{i+1} - \mathbb{E}[v_{i+1}], v_{i+1}^2 - \mathbb{E}[v_{i+1}^2], v_{i+1}^3 - \mathbb{E}[v_{i+1}^3], v_{i+1}^4 - \mathbb{E}[v_{i+1}^4]]^\top$. This corresponds to setting $pow = 4$ and $cpo = 0$.[3]

[2] https://github.com/mrockinger/HigherMoments.

[3] For faster estimation, we generated the explicit expression of the objective function using a symbolic language (Mathematica).

`gmmW.c2.p3.l2` generates the minimum-distance estimator where we set the power $pow = 3$, the copower $cop = 2$, and where we allow for up to 2 lags. All moments combined, this estimation will involve 12 moment conditions.

`gmmW.c3.p4.l6` generates the minimum-distance estimator where $pow = 4$, $cop = 2$, and where we allow for up to 6 lags. This method will involve 46 moment conditions.

In his book Iacus (2008) discusses how one can exploit orthogonality conditions resulting from a linearization of the CIR process. This yields the following moment conditions.

For model `gmmW.1.c`, we consider conditional moments and their orthogonality conditions

$$\begin{aligned} u_i^{gmm.1.c}(\kappa,\sigma) = \big[& v_{i+1} - \mathbb{E}\left[v_{i+1}|\mathcal{F}_i\right], v_i\left(v_{i+1} - \mathbb{E}\left[v_{i+1}|\mathcal{F}_i\right]\right), \\ & \mathbb{V}\left[v_{i+1}|\mathcal{F}_i\right] - \left(v_{i+1} - \mathbb{E}\left[v_{i+1}|\mathcal{F}_i\right]\right)^2, \\ & v_i\left\{\mathbb{V}\left[v_{i+1}|\mathcal{F}_i\right] - \left(v_{i+1} - \mathbb{E}\left[v_{i+1}|\mathcal{F}_i\right]\right)^2\right\}\big]^\top. \end{aligned}$$

The model `gmmW.2.c` augments the previous moment conditions by third moments. We can write explicitly

$$u_i^{gmm.2.c}(\kappa,\sigma) = \left[u_i^{gmm.1.c}, v_{i+1}^3 - \mathbb{E}\left[v_{i+1}^3|\mathcal{F}_i\right], v_i\left(v_{i+1}^3 - \mathbb{E}\left[v_{i+1}^3|\mathcal{F}_i\right]\right)\right]^\top$$

Using the same logic model `gmmW.3.c` increments the previous model by fourth-order moments. This can be written as

$$u_i^{gmm.3.c}(\kappa,\sigma) = \left[u_i^{gmm.2.c}, v_{i+1}^4 - \mathbb{E}\left[v_{i+1}^4|\mathcal{F}_i\right], v_i\left(v_{i+1}^4 - \mathbb{E}\left[v_{i+1}^4|\mathcal{F}_i\right]\right)\right]^\top.$$

So far, we have been dealing with conditional moments. It is also possible to use similar orthogonality conditions generated from unconditional moments. This yields the first model `gmmW.1.uc` with moments as

$$\begin{aligned} u_i^{gmm.1.uc}(\kappa,\sigma) = \big[& v_{i+1} - \mathbb{E}\left[v_{i+1}\right], v_i v_{i+1} - \mathbb{E}\left[v_i v_{i+1}\right], \\ & \mathbb{V}\left[v_{i+1}\right] - \left(v_{i+1} - \mathbb{E}\left[v_{i+1}\right]\right)^2, \\ & \left(\mathbb{E}\left[(v_i v_{i+1})^2\right] - \mathbb{E}\left[v_i v_{i+1}\right]^2\right) - \left(v_i v_{i+1} - \mathbb{E}\left[v_i v_{i+1}\right]\right)^2\big]. \end{aligned}$$

We can again extend this model by using third-order moments. Explicitly this gives model `gmmW.2.uc` with as moments

$$u_i^{gmm.2.uc}(\kappa,\sigma) = \left[u_i^{gmm.1.uc}, v_{i+1}^3 - \mathbb{E}\left[v_{i+1}^3\right], v_i v_{i+1}^3 - \mathbb{E}\left[v_i v_{i+1}^3\right]\right]^\top$$

As previously, we can extend the above by introducing fourth-order moments. This gives us a model `gmmW.3.uc` written as

$$u_i^{gmm.1.uc}(\kappa,\sigma) = \left[u_i^{gmm.2.uc}, v_{i+1}^4 - \mathbb{E}\left[v_{i+1}^4\right], v_i v_{i+1}^4 - \mathbb{E}\left[v_i v_{i+1}^4\right]\right]^\top.$$

With those various moment in mind, it is possible to generate moment conditions

$$g_M(\kappa,\sigma) = \frac{1}{M}\sum_{i=1}^{M} u_i^{model}(\kappa,\sigma).$$

Eventually, the parameter can be estimated using the objective function (6).

This concludes our list of moment-based estimates. Let us now turn to maximum likelihood methods.

4.3 ML and pseudo-ML approaches

In the literature, several ML-type methods are available ranging from exact to some approximation. `ml.ex` implements the exact likelihood. As already discussed in Zhou (2001); Bollerslev and Zhou (2002); Andersen (2008) the evaluation of the exact likelihood requires evaluation of the noncentral chi-squared density. This in turn requires evaluating a Bessel function, and a slow iterative process achieves this. In many cases, this iteration fails to converge correctly.

As an alternative, Pseudo-ML approaches have been developed. Here, the true transition density is replaced by another, simpler density. Mostly, this replacement is based on some specific discretization of (3). For instance a transition density can be obtained using an Euler discretization. Having two low-frequency points in time, say s and t, with $s \le t$, let $\Delta_{LF} = t - s$, then after using a direct Euler approximation of (3) one obtains, under the assumption of Gaussian innovations, the conditional distribution

$$\tilde{v}_t|\tilde{v}_s \sim \mathrm{N}\left(\tilde{v}_s + \kappa\left(\theta - \tilde{v}_s\right)\Delta_{LF}, \sigma^2\tilde{v}_s\Delta_{LF}\right). \tag{7}$$

From this point onward, a ML procedure can be easily implemented; essentially it boils down to OLS. We denote by `ml.eu` the Peusdo-ML approach based on an Euler discretization of the original variance process.

A different branch of Pseudo-ML approaches is generated by local linearization. Shoji and Ozaki (1998) linearize the drift term of (3). A justification of their method and indications on how to implement it can be found in their paper and will not be repeated here. Connected to these Pseudo-ML ideas, Pedersen (1995) presented a simulated or in-fill likelihood method. This approach will make the approximation error to the original distribution, which is naturally arising from any discretization, arbitrarily small. This is done by isntroducing simulated observations between the observed points in time. The efficiency of this method was discussed and further improved by Durham and Gallant (2002). Additionally, several Quasi-ML approaches are available. The difference between Pseudo- and Quasi-ML approaches is that in the

latter, the target transition density is approximated rather than replaced by a closed-form density based on some specific discretization approach to the original diffusion. For example, Kessler (1997) uses an Ito-Taylor expansion while Aït-Sahalia (2002) uses a Hermite expansion.

`ml.sh` will denote the estimates resulting from using the pseudo-ML method proposed by Shoji and Ozaki (1998).

Since it is well-known that for small sample sizes, all of the discussed ML-related methods tend to overestimate the value of κ, Phillips and Yu (2009) introduced a Jackknife procedure to improve those estimates.[4] Formally, let M be as above the total number of filtered variance observations. We can split this sample into L consecutive subsamples with ℓ observations, that is $M = \lfloor L \times \ell \rfloor$. The jackknife estimator is then defined as

$$\hat{\psi}_{jack} = \frac{L}{L-1}\hat{\psi}_M - \frac{\sum_{i=1}^{L}\hat{\psi}_{\ell i}}{L^2 - L},$$

where ψ denotes the parameter vector (κ, θ, σ), $\hat{\psi}_M$ the corresponding ML-type estimator based on the complete sample and $\hat{\psi}_{\ell i}$ the respectively chosen ML-type estimator based on the i-th subsample of size ℓ. In a preliminary investigation we found that $L = \lfloor M^{1/7} \rfloor$ seems to be an appropriate choice for partitioning the original sample. In the case where one estimates the parameters $\hat{\psi}_M$ and $\hat{\psi}_{\ell,i}$ under constraints, it is possible that the jackknife estimator provides an estimate beyond the constraints. This is not astonishing since the goal of the jackknife method is to project parameters into a region which is otherwise difficult to reach.

Since the jackknife method can be applied to various settings, we introduce the following acronyms: `jk.ex` is the jackknife procedure applied to the exact ML method, `ml.ex`. `jk.eu` is the jackknife method applied to the `ml.eu` approach, which uses the Euler discretization. And `jk.sh` finally, is the jackknife applied to the method proposed by Shoji and Ozaki (1998) and therefore extends `ml.sh`.

Now that the estimation methods are defined, let us discuss the overall simulation setting.

5 Simulation framework

5.1 Choice of parameters and choice of simulation algorithm

In Table 1 we present, for the volatility dynamic parameters θ, κ, and σ, estimates that can be found in the literature. We also present the correlation parameter for cases where the estimation is for the entire Heston model. The caption of the table provides the references to where the parameters can be found.

We notice that the Feller ratio ranges from 0.02 to 653 and therefore covers an extensive range of values. Out of the ten models presented in this table, five do not satisfy the Feller condition. This important observation implies that research related to the CIR process, be it theoretical, or about the discretization of this process, its

[4] This issue is the same as estimating the autoregressive parameter of an AR(1) process, in particular when the parameter is close to one (meaning that the AR(1) process is close to being an integrated, I(1), process).

Table 1 Various parameter constellations found in the literature for various assets

	Model									
par	1	2	3	4	5	6	7	8	9	10
κ	0.500	0.300	1.000	6.210	2.000	5.300	0.013	0.026	0.200	0.240
θ	4%	5%	9%	2%	9%	24%	53%	2%	6%	3%
σ	1	0.9	1	0.019	1	0.38	0.0734	0.2155	0.0415	0.0269
ρ	−0.9	−0.5	−0.3	−0.7	−0.3	−0.57	−0.58	−0.268	–	–
$\frac{2\kappa\theta}{\sigma^2}$	0.04	0.04	0.18	653.68	0.36	17.76	2.56	0.02	13.47	16.58

Models 1–3 are as in Andersen (2008), models 4–5 are as in Duffie et al. (2000). For stocks, model 6 is found in Pan (1999). Model 7 is in Andersen et al. (2002) for daily SP data, and model 8 is for daily NASDAQ data and is obtained by Eraker et al. (2003). Estimates for models 9 and 10 are in Pan (1999) for weekly interest rates and dividend yields. Here, the volatility dynamic (CIR process) is the focus instead of log-returns. The ratio $2\kappa\theta/\sigma^2$ defines the Feller ratio; values below 1 indicate that a process may come close to 0 and stay there for significant time

simulation, and its estimation should all encompass the case where the Feller condition is not satisfied.

We observe that a κ as small as 0.013 for SP500 returns implies a half life of $\log(2)/\kappa$ of 53 days; thus that the volatility dynamic will be very persistent and that for stock market data, by analogy with AR processes close to the unit root case, one may expect in general difficulties in the estimations due to a bias.

Based on the values in Table 1 and since this parameter can be estimated easily just using a mean (since it represents the long-run volatility), we decided to use two levels of θ, namely 0.4 and 2. We select $\kappa \in \{0.5, 3.5, 7\}$ and finally, pick $\sigma \in \{0.1, 2.05, 4\}$. This allows for a collection of 18 combinations.

Let us now define the setup of the Monte Carlo study. There is a large literature on how to simulate the Heston process. The methods range from simple Euler-type linearizations to sophisticated exact schemes, see Kahl and Jäckel (2006); Broadie and Kaya (2006); Lord et al. (2010); Glasserman and Kim (2011); Zhu (2011), and more recently (Okhrin et al. 2022) which contains further references.

The difficulty in all those situations is to find a compromise between exact sampling as in Broadie and Kaya (2006), which is, however, rather slow due to the necessity to simulate from a Gamma distribution, and discrete approximations like the Euler scheme, which is fast but less accurate, if the time step is relatively large. In the following, we use the scheme proposed by Andersen (2008) with the correction for skewness by Okhrin et al. (2022), which seems to be a good compromise, being fast and accurate at the same time. An implementation may be found in a GitHub depository.[5]

5.2 Spot variance estimation quality

The first step in our empirical investigation is to discuss the estimation of spot variance. To do so, we start our investigations by examining the quality of the estimates of (5) using four different kernels. Those are a uniform kernel, as suggested in ART, as well

[5] https://github.com/mrockinger/CIR-Heston-Simulation.

Table 2 RMSE for the spot variance. We simulate 500 time-series of length 99,990 and obtain the RMSE along each trajectory of the estimated spot variance and the actual variance

		$J = 100$, $J_{gap} = 10$						$J = 16$, $J_{gap} = 96$					
		$\kappa = 0.5$		$\kappa = 3.5$		$\kappa = 7$		$\kappa = 0.5$		$\kappa = 3.75$		$\kappa = 7$	
σ	$\theta =$	0.4	2	0.4	2	0.4	2	0.4	2	0.4	2	0.4	2
0.1	U	**0.09**	**0.29**	**0.07**	**0.29**	**0.07**	**0.29**	**0.21**	**0.73**	**0.16**	**0.71**	**0.17**	**0.71**
	G	0.11	0.38	0.09	0.37	0.08	0.37	0.28	0.98	0.21	0.96	0.22	0.96
	E	0.09	0.32	0.07	0.31	0.07	0.31	0.24	0.83	0.18	0.81	0.19	0.81
	OS	0.18	0.63	0.13	0.62	0.13	0.62	0.45	1.53	0.33	1.50	0.32	1.49
2.05	U	**0.37**	**0.77**	**0.27**	**0.64**	0.25	**0.60**	**0.66**	**1.54**	**0.44**	**1.15**	**0.40**	**1.08**
	G	0.43	0.91	0.30	0.74	0.28	0.70	0.80	1.93	0.49	1.36	0.44	1.28
	E	0.39	0.82	0.28	0.68	0.26	0.64	0.72	1.70	0.46	1.23	0.42	1.16
	OS	0.53	1.15	0.26	0.79	**0.24**	0.76	1.15	2.76	0.61	1.88	0.54	1.78
4	U	**0.70**	**1.44**	**0.51**	**1.15**	0.47	1.07	**1.30**	**2.78**	**0.82**	**1.90**	**0.73**	**1.73**
	G	0.83	1.69	0.57	1.29	0.53	1.21	1.57	3.44	0.89	2.11	0.78	1.92
	E	0.75	1.54	0.54	1.22	0.50	1.14	1.41	3.06	0.85	1.98	0.76	1.81
	OS	1.02	2.03	0.46	1.16	**0.41**	**1.06**	2.27	4.77	1.06	2.68	0.90	2.38

We present the average of those RMSE. Letters U, G, E, and OS represent the following kernels: Uniform, Gaussian, Epanechnikov, and One-Sided. The number of observations in the estimation windows is J. Each observation window is separate from the other by J_{gap} observations. Bold font indicates the best average estimate (the ranking involves all decimals)

as a Gaussian, Epanechnikov, and a One-Sided kernel. In the remainder, we will refer to these weighting schemes as U, G, E, and OS.

For this comparison, we simulate from a general high-frequency sampling scheme setting $\Delta_{HF} = 0.001$, which refers to roughly four observations per trading day. Further, we will simulate a sample size of $n = 99,990$, which is approximately 100 years of data. Such a large sample is required since we also want to discuss the consequence of bucket size and distance between buckets as discussed in Sect. 3.

On the one hand, inspired by ART, we will set $J = 100$ and $J_{gap} = 10$ which results in $M = 909$ and $\Delta_{LF} = 0.11$. This choice of J and J_{gap} corresponds to one estimated variance observation per month and then waiting 2.5 days. Note that in this aggregation setup, we are at the upper bound of an optimal J, which naturally causes a relatively large Δ_{LF}. Nevertheless, this is still near the optimal lower bound of ξ.

On the other side, corresponding to Bandi and Renò (2018), we will set $J = 16$ and $J_{gap} = 94$. This corresponds to an estimation of volatility by sampling the variance process for four days and then waiting for 23.5 days.

We report the average of the RMSE of the point-wise estimators. The RMSE is the square root of the (L^2) mean squared distance between the filtered and the actual variance for every low-frequency point in time for each of the 500 samples and the 18 process parameter combinations. The results can be found in Table 2. Overall, one can see that the RMSEs are naturally bigger for larger values of θ, the long-run mean. Notice that the ratio of RMSE over θ also increases demonstrating that this increase is

a genuine phenomenon. Also, the RMSE seems to be an increasing function of σ and a moderate decreasing function of κ. Aggregation-wise, it seems that more significant errors occur for the second setup, that is, when J is chosen smaller than J_{gap}.

In terms of the kernel, which represents an important choice for the following estimations, the uniform kernel of ART dominates the other approaches except for five cases out of 36, where the one-sided kernel leads to a smaller RMSE. Having these results at hand, the overall conclusion is that the approach of ART, in terms of averaging, kernel-wise as well as in terms of J and J_{gap}, delivers the best estimation results. In the following, building on the conclusion of this preliminary investigation, we will be using the uniform kernel to estimate spot volatility.

5.3 Analysis of the simulation

A preliminary remark is that we do not report results for the estimation of the mean level θ, since this estimator was nearly unbiased for all considered parameter combinations with only minor RMSEs.

We also try two ways of determining the starting values for these procedures: In the first case, we set the starting values equal to the parameters estimated by the method of ART. In the second case, we draw starting values for κ, σ as $U(0.01, 10)$ that is as a uniform distribution with a range of 0.01 up to 10. For each set of parameters, we draw 10,000 starting values and perform for each set the estimation. Eventually, we retain the ten best estimates based on the value of the objective function, for which we run a final BFGS optimization routine. The objective of this procedure is to capture as many local minima as possible and then retain as a final estimate the overall minima. We will refer to the latter procedure by suffixing an '`.rs`' (for random starting value) to the original method's abbreviation.

At this stage, we wish to discuss the results of the estimation on simulated data. All the estimations were done in `R`, with the optimizer set to use the BFGS algorithm. The lower boundaries of the parameters were set to 0.01 and the upper boundaries to 10. Together we have 18 parameter sets as stated in Sect. 5.1, there are 19 basic estimation methods, and in addition, we experimented with random starting values.

Some methods, like ART, deliver directly parameters, by construction. Thus, for this model, the estimator is not based on a numerical optimization and since there is no starting value, we also do not need to experiment with random starting values.

Other methods like the jackknife combine estimations from sub-samples and therefore the choice of starting values is less relevant. Altogether, this generates some $18 \times 32 = 576$ models. We also ran those models either on spot variance processes or the corresponding true volatility, sampled on the low frequency. This doubles the model estimations. Each model simulation involves 500 samples with $n = 99,990$ observations. The resulting estimation study is, therefore, a demanding task, for which we used a dedicated server with 48 cores.

Our programs return for each of the models the root mean square error (RMSE) and a count of the number of times an estimation succeeded. We consider that an estimation succeeded when all the parameters remained within their respective boundaries. The analysis of those 576×2 estimations can therefore be viewed as a data analysis that

we are going to discuss presently.[6] We structure our discussion as follows. First, we will inspect graphically the objective functions obtained from a random draw. Next, we present for all models the number of times we had a convergence out of these 500 simulations, and will eventually select all those models where the number of conversions exceeds a certain threshold.

Once we have selected models that converge sufficiently often, we will ask if for certain parameter constellations, the parameter estimates are better or worse than for others. The intuition why certain parameter constellations could be worse than others comes from the fact that in some instances the Feller condition does not hold. If this is the case, we have observed graphically that volatility can come close to zero and remain there for a long time. Then, one fits a model on a series of data containing lots of zeros resulting in poor estimates.

Similarly, we can neglect the parameters and focus on the overall RMSE for a given model. This will allow us to rank the various models and to select, at least in theory, the best model.

As discussed, let us start by graphically comparing the objective functions for a random draw.

5.3.1 Inspection of the objective function

In Figure 1, we display the profiles of the objective function of several GMM and ML methods. We have chosen $\theta = 0.4$, κ takes as value 0.5 and σ is set to 4. This corresponds to a situation where the Feller ratio equals 0.15 and, therefore, where the Feller condition is not satisfied. The sample is purely random, and the presented figures reflect a scenario where the estimation complexity is initially unknown.

Those plots contain the true parameters as a big red dot. The blue diamond is the ART estimation. The red star denotes the estimation when we use perform several thousand estimations with random initial parameters. The legend in the figure presents other methods.

As those plots reveal, the minimum of the objective function can lay relatively far away from the true values. This is due to the finite sample size of the simulation. Next, we inspect the shape of the objective functions. If we focus on `gmmW.c2.p3.l2` we notice that the objective function can take elongated banana-shaped forms. This is reminiscent of the Rosenbrock function, a textbook case of a difficult-to-optimize function. Similarly, when we switch to the ML-based methods, we notice that the log-likelihood function can be either extremely flat as it is for the exact ML method. For the Euler or local-linearized (Shoji and Ozaki 1998) type methods, the log-likelihood maximum is somehow squished to the domain's boundary. Thus, given the shape of the objective function, there is no chance for even the best opimizers to converge to the true values.

In this figure, we also present the estimates obtained by the deterministic `ART` model. We also perform a grid-based research of the optimum instead of a numerical optimization. As the estimates demonstrate, we are confronted with a rather complicated situation. Given the shape of the objective function, it will be a daunting task

[6] In an online Appendix the interested reader can find the outputs from those runs.

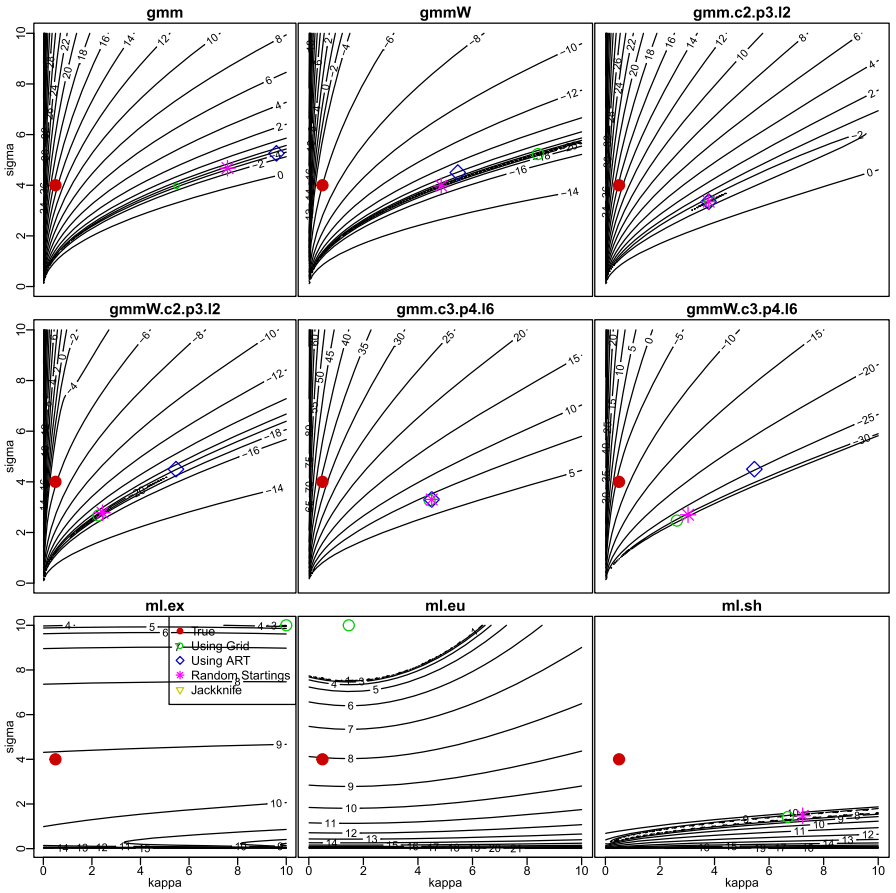

Fig. 1 Objective functions for the estimation methods given the parameters $\kappa = 0.5$, $\sigma = 4.0$ and $\theta = 0.4$. In this setting, the Feller condition is not satisfied. The simulated series are of length 99,990. The red dot is the true value. We notice huge differences in the shapes of the objective functions

to find a satisfactory estimate whatever the estimation method used. This observation will follow us along in the remaining discussion of this paper.

Let us presently turn to discussing the convergence properties of the various models.

5.3.2 Methods retained based on convergence

In the following analysis, we select a given set of parameters (κ, θ, σ), perform 500 simulations, and count the number of converged simulations. In Table 3, we display on the left side the results of the estimations when we use as starting values of the estimations the parameter estimates produced by ART's method, and in the right part when we use the '`.rs`' scheme described above.[7]

[7] In some rare cases when `ART` failed to generate finite numbers, we set arbitrarily $\kappa = 2$ and $\sigma = 1$ as starting values.

Table 3 We present the selected estimation methods along with their acronyms and provide statistics on the number of convergences achieved during the numerical optimizations

fix starting values				random starting values			
model	min	mean	max	model	min	mean	max
gmmW	276	488	500	gmm.rs	500	500	500
gmm	299	486	500	gmmW.rs	500	500	500
ml.sh	0	471	500	ml.sh.rs	0	476	500
art	2	471	500	gmm.c2.p3.l2.rs	0	476	500
gmmW.2.c	212	470	500	gmmW.2.c.rs	226	471	500
gmmW.3.c	106	468	500	gmmW.3.c.rs	182	469	500
gmm.c3.p4.l6	110	466	500	gmmW.1.uc.rs	48	459	500
gmm.c2.p3.l2	0	465	500	gmmW.1.c.rs	21	457	500
gmmW.1.c	21	458	500	gmmW.c2.p3.l2.rs	0	446	500
gmmW.c3.p4.l6	0	457	500	ml.eu.rs	0	443	500
jk.sh	0	457	500	gmmW.3.uc.rs	116	426	500
gmmW.1.uc	18	456	500	gmmW.2.uc.rs	116	410	500
gmmW.c2.p3.l2	0	451	500	ml.ex.rs	0	19	482
ml.eu	0	443	500				
jk.eu	0	434	500				
gmmW.3.uc	117	421	500				
gmmW.2.uc	115	408	500				
jk.ex	0	15	252				
ml.ex	0	14	249				

In the left part of the table, we initialize the parameters using parameter estimates resulting from Azencott et al. (2020). In the right part, we estimate the model with random starting values and retain the best fitting model. Heston's model is simulated 500 times for a given set of parameters. A set of parameter estimates results for each of those 500 simulations unless the algorithm hits boundaries or does not converge after a predetermined amount of iterations. Min, average, and max indicate the number of minimal, average, and maximal number of convergencies across all sets of parameters. We consider that an estimation converged if the resulting parameters did not hit the upper or lower boundaries which needed to be incorporated in the numerical optimizations to ensure convergence

If we focus on the left part, the method resulting in the highest average convergences is gmmW (across all parameter combinations) closely followed by gmm. For gmmW, we notice that across all parameter sets, in the worst constellation, 276 estimations converged. On average, 489 converged. For some parameters, convergence occurred in all 500 simulations. It is interesting to notice that adding more lags and thus information to the GMM, such as in gmm.c2.p3.l2 does not necessarily lead to better estimates. Adding complexity thus does not seem to help to get good estimates. It seems to help to use the structure of a process as much as possible, for instance, measuring how two consecutive observations are linked via their conditional distribution. This can explain the relative success of the gmmW.2.c and gmmW.3.c methods, which use conditional information instead of unconditional one.

At the bottom of the Table 3, we find the exact ML estimation, `ml.ex`. Since this method has difficulties converging, see Zhou (2001); Bollerslev and Zhou (2002), it does not come as a surprise that the jackknife method `jk.ex`, which is a portfolio of ML estimations, fails also. For the ML-based methods, one can conclude that `ml.sh` seems preferable to `ml.eu`. Those findings align with the literature, see, for example, Durham and Gallant (2002).

When we turn to the RHS of the Table 3, we notice that for certain methods at the top of the table, the use of random starting values permits obtaining more convergences than in the deterministic starting-value case. For the models in the lower part of the table, using random starting values improves the number of convergences only marginally if at all.

At this stage, we decided to reduce the number of models by eliminating those where across all parameter values, there were fewer than 90% successful estimations on average, out of the 500 possible ones. This led to the elimination of 6 models: `gmmW.2uc`, `gmmW.3uc`, `ml.ex`, `jk.ex`, `ml.eu`, `jk.eu`.

By eliminating those models, we obtain parameter estimates that are not on the boundaries, and, therefore, we can investigate the parameters themselves. One obvious question is if all parameter sets can be estimated with the same ease or if there are settings where it is more difficult to find correct estimates.

5.3.3 Are some parameters more complicated to estimate?

Across estimation methods, we address the issue if there are constellations of parameters where the estimation is more complicated? This investigation is important since it demonstrates that it will be more challenging to estimate some or all the parameters for some parameter combinations, and this will not depend on the method used.

To formally investigate this issue, we regress the root mean squared error of the estimations on the parameters as in:

$$RMSE(\hat{x})_i = \beta_0 + \beta_\theta \theta_i + \beta_\kappa \kappa_i + \beta_\sigma \sigma_i + \beta_{VolT} \mathbb{I}_{VolT,i} + \beta_{Feller} \log(Feller)_i + u_i.$$

Here, i runs over all methods and all parameter constellations.[8] $\mathbb{I}_{VolT}$ is a dummy variable indicating if we estimate the parameters on the actual variances, in which case it takes the value 1, or if we use spot volatility in which case its value will be 0. Thus, $VolT$ denotes the volatility type.

The estimates of this model are found in Table 4. We first notice that the parameter κ and σ tend to be less well estimated than θ. Intuitively, if κ is underestimated, meaning that the process is closer to being integrated than it theoretically is, the volatility will also be large. The parameter β_{VolT} in that table is slightly negative but insignificant. This important result demonstrates that if it is difficult to obtain reasonable estimations, it is not because of the bad estimation of volatility but because the available estimation methods have difficulties finding the correct parameters.

We notice that β_θ is insignificant for the estimation of θ and κ. However, higher levels of θ yield better estimates of the σ, the variability of the variance process.

[8] For this exercise we used the deterministic starting values.

Table 4 We consider the RMSE for the parameters and investigate if across all estimation methods some parameters are estimated with a larger error and if so if this depends on one of the underlying parameters

param	$RMSE(\hat{\theta})$	$RMSE(\hat{\kappa})$	$RMSE(\hat{\sigma})$	$RMSE(\hat{\theta})$	$RMSE(\hat{\kappa})$	$RMSE(\hat{\sigma})$
β_θ	0.084	−0.079	−0.072	0.104	−0.194	−0.049
	(0.045)*	(0.0489)	(0.025)**	(0.054)*	(0.059)***	(0.030)*
β_κ	−0.067	0.019	−0.087	−0.059	−0.028	−0.078
	(0.014)***	(0.015)	(0.007)***	(0.018)***	(0.020)	(0.010)***
β_σ	0.076	0.098	0.188	0.039	0.307	0.141
	(0.022)**	(0.024)***	(0.012)***	(0.058)	(0.065)***	(0.033)***
β_{VolT}	0.004	−0.059	−0.018	−0.002	−0.059	−0.018
	(0.072)	(0.078)	(0.039)	(0.072)	(0.077)	(0.039)
$\beta_{\log\ Feller}$				−0.020	0.114	−0.023
				(0.030)	(0.033)***	(0.017)
Const	0.185	1.077	−0.500	0.240	0.781	0.559
	(0.099)*	(0.108)***	(0.054)***	(0.128)	(0.137)***	(0.069)***
$\bar{R}^2$	0.488	0.042	0.455	0.479	0.066	0.455

VolT is a dummy variable taking the value zero if the parameter has been estimated on the spot volatility and one if it is the known actual volatility. *, **, and *** mark significance levels of 10%, 5%, and 1%, respectively

Intuitively, if θ is large, the process will evolve far from zero, and it will be more difficult for the process to come to the neighborhood of zero, which might result in a pathological trajectory. The parameter β_κ is slightly significant for the RMSE of $\hat{\theta}$ and $\hat{\sigma}$ but not of its value κ. A larger value of κ represents faster mean reversion of the process, implying that the processes oscillate more around the mean and, thus, θ should be better estimated. With a better mean estimation, the volatility of volatility should also be better estimated. Those are all well-known features from estimating processes close to $I(1)$.

The parameter β_σ is always positive and highly significant for $\hat{\kappa}$ and $\hat{\sigma}$, except for $\hat{\theta}$ where it is significantly only at the 5% level. More variability in the dynamic of variance, because σ is larger, will translate into a worse estimation of all the parameters. The intuition is that for larger σ, the process comes closer to 0, in which case the trajectory may become pathological and estimating the parameters more difficult.

In the following three columns of Table 4, we introduce in addition the Feller ratio. As the coefficient β_{Feller} shows, this addition does not matter for the estimations beyond the information already provided by the parameters. Only for $\hat{\kappa}$ does a larger Feller ratio lead to a larger RMSE. The coefficient of determination remains, however, very small.

Having concluded this investigation, we will investigate the best functioning model regarding overall RMSE.

5.3.4 Horse-race of models

Table 5 presents the various remaining estimation methods and indicates the overall RMSE. In the left part of each pair of columns, we indicate the value of the RMSE. In the right part, we show the rank of the model. Rank 1 is the model with the overall smallest RMSE. The columns labeled 'actual vol' are less representative of a real-life application since the true variance process is unknown.

We notice that the jackknife method of Shoji and Ozaki (1998), namely `jk.sh` is ranked first in terms of smallest overall RMSE, followed by the maximum likelihood estimate `ml.sh`. However, on actual variance data, `ml.sh` would seem less successful. Concerning the GMM methods, no clear winner emerges. In their simulation (Bollerslev and Zhou 2002) report very good estimates for GMM estimates. Their processes satisfy the Feller condition and may be easier to estimate.[9]

While being very quick, the method by ART is also on the podium with a third place for the estimation of κ. It is 6th out of 12 for σ. We also notice that the best RMSE for κ is 0.92, which is about three times larger than the best estimate of σ, emphasizing the difficulty of measuring this parameter correctly.

As we could expect from our discussion of the objective functions in Figure 1, it is generally challenging to estimate the parameters of the volatility process. No method uniformly dominates all the other methods. So, which model should one retain? Model `art` is the fastest since it does not require numerical optimization. Its average RMSE remains similar to the ones obtained by the best model, namely 1.01 versus 0.92 for κ

[9] Their scenarios are $\{\kappa, \theta, \sigma\}$ taking the sets of values (0.03, 0.25, 0.1), (0.1, 0.25, 0.1), and (0.1, 0.25, 0.2) with associated Feller ratio 1.5, 5, and 1.25.

Table 5 Which model performs best? We split the estimations according to κ and σ and the estimation method

	κ				σ			
	spot vol.		actual vol.		spot vol.		actual vol.	
method	average RMSE	rank est	average RMSE	rank est	average RMSE	rank est	average RMSE	rank est
art	1.01	**3**	1.33	8	0.41	6	0.34	**3**
gmm	1.38	9	1.33	9	0.45	7	0.38	6
gmm.c2.p3.l2	1.27	7	1.21	7	0.47	9	0.40	8
gmm.c3.p4.l6	1.52	11	0.89	**3**	0.33	5	0.60	11
gmmW	1.11	4	1.15	6	0.46	8	0.43	9
gmmW.1c	1.18	6	1.13	5	0.63	11	0.39	7
gmmW.1uc	1.75	12	0.83	**2**	0.54	10	0.80	12
gmmW.2c	1.14	5	1.65	12	0.31	**3**	0.34	4
gmmW.3c	1.39	10	1.40	10	0.86	12	0.36	5
gmmW.c3.p4.l6	1.34	8	1.01	4	0.27	**1**	0.31	**2**
jk.sh	0.92	**1**	0.82	**1**	0.33	4	0.25	**1**
ml.sh	0.92	**2**	1.47	11	0.28	**2**	0.52	10

We then present the average RMSE of the parameter estimates for all the models investigated. We also stratify the estimation results based on whether actual volatility or estimated spot volatility was used. Each group of estimates gets ranked. For instance for the estimation of κ on estimated spot volatility, jk.sh produced the smallest RMSE and gets therefore the rank 1. The models are sorted according to alphabetical order

and 0.41 versus 0.28 for σ. It is for those reasons that we use model `art` to generate starting values for the estimations involving iterative methods (besides the involved '.rs' random starting value approach).

At this stage, we have discussed the difficulties in estimating the parameters from the variance process. It is time to discuss the estimation of the remaining parameters of the model. There is the mean of the returns and the correlation between the innovations, ρ. The mean is easy to obtain, with the usual caveat that means of log-returns tend to be unstable, but the estimation of correlation is more involved.

6 Conclusion

In this paper, we model and sample spot variance by combining two dominant strategies of the literature: The kernel-based method introduced by Kanaya and Kristensen (2016) and the bucket approach presented in Bandi and Renò (2018) and Azencott et al. (2020) (ART). First, we perform a broad Monte Carlo study, investigating the impact of the aggregation level on the estimator on the one hand and the goodness of four different kernels on the other side. Overall, we find that the aggregation level should be chosen as high as possible, thus one should use as many high-frequency observations given in

one low-frequent time interval as possible. Weighting-wise, we found that the uniform kernel of ART performs best.

Next, we discuss several ML-, MM- and GMM-type parameter estimation approaches within the framework of Heston's stochastic volatility model. We propose a novel GMM-type estimator involving moment conditions that utilize higher (un)conditional (co)moments of the spot volatility. In a second Monte Carlo study, we analyze the performance of those estimation methods on actual variance realizations and on spot variance estimates. For a situation where the characteristics of the data are not known, either the method of ART or of Shoji and Ozaki (1998), possibly in a jackknifed version should be best. We can also conclude that using higher moments in the general GMM procedure does not give better estimation results in most cases but increases the estimators' RMSE.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Azencott R, Ren P, Timofeyev I (2020) Realised volatility and parametric estimation of Heston SDEs. Financ Stochast 24:723–755

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

Cox JC, Ingersoll JE, Ross SA (1985) A theory of the term structure of interest rates. Econometrica 53(2):385–408

Andersen T, Bollerslev T, Diebold F, Ebens H (2001) The distribution of realized stock volatility. J Financ Econ 61:43–76

Barndorff-Nielsen OE, Shephard N (2002) Econometric analysis of realized volatility and its use in estimating stochastic volatility models. J R Stat Soc Ser B Stat Methodol 64(2):253–280

Andersen TG, Lund J (1997) Estimating continuous-time stochastic volatility models of the short-term interest rate. J Econom 77(2):343–377

Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. J Econom 108(2):281–316

Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and forecasting realized volatility. Econometrica 71(2):579–625

Bollerslev T, Zhou H (2002) Estimating stochastic volatility diffusion using conditional moments of integrated volatility. J Econom 109(1):33–65

Renò R (2006) Nonparametric estimation of stochastic volatility models. Econ Lett 90(3):390–395

Kanaya S, Kristensen D (2016) Estimation of stochastic volatility models by nonparametric filtering. Econometr Theor 32(4):861–916

Bandi FM, Renò R (2016) Price and volatility co-jumps. J Financ Econ 119(1):107–146

Bandi FM, Renò R (2018) Nonparametric stochastic volatility. Econometr Theor 34(6):1207–1255

Shoji I, Ozaki T (1998) Estimation for nonlinear stochastic differential equations by a local linearization method. Stoch Anal Appl 16(4):733–752

Eraker B, Johannes M, Polson N (2003) The impact of jumps in equity index volatility and returns. J Financ 58(3):1269–1300
Jacquier E, Polson NG, Rossi PE (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. Journal of Econometrics 122(1):185–212
Johannes M, Polson N (2010) MCMC methods for continuous-time financial econometrics. In: Aït-Sahalia Y, Hansen LP (eds) Handbook of financial econometrics: applications. Elsevier, Amsterdam, pp 1–72
Bates DS (1996) Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. Rev Financ Stud 9(1):69–107
Duffie D, Pan J, Singleton K (2000) Transform analysis and asset pricing for affine jump-diffusions. Econometrica 68(6):1343–1376
Feller W (1951) Two singular diffusion problems. Ann Math 54:173–182
Okhrin O, Rockinger M, Schmid M (2022) Distributional properties of continuous time processes: from CIR to Bates. AStA Adv Stat Anal 1–23
Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50(4):1029–1054
Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3):703–708
Newey WK, West KD (1994) Automatic lag selection in covariance matrix estimation. Rev Econ Stud 61(4):631–653
Iacus SM (2008) Simulation and inference for stochastic differential equations: with R examples, vol 486. Springer, New York
Zhou H (2001) Finite sample properties of EMM, GMM, QMLE and MLE for a square-root interest rate diffusion model. J Comput Financ 5(2):89–122
Andersen L (2008) Simple and efficient simulation of the Heston stochastic volatility model. J Comput Financ 11(3):1–43
Pedersen AR (1995) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scand J Stat 55–71
Durham GB, Gallant AR (2002) Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. J Bus Econ Stat 20(3):297–338
Kessler M (1997) Estimation of an ergodic diffusion from discrete observations. Scand J Stat 24(2):211–229
Aït-Sahalia Y (2002) Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. Econometrica 70(1):223–262
Phillips PC, Yu J (2009) Maximum likelihood and Gaussian estimation of continuous time models in finance. In: Mikosch T, Kreiß JP, Davis R, Andersen T (eds) Handbook of financial time series. Springer, Berlin, Heidelberg, pp 497–530
Kahl C, Jäckel P (2006) Fast strong approximation Monte Carlo schemes for stochastic volatility models. Quant Financ 6(6):513–536
Broadie M, Kaya Ö (2006) Exact simulation of stochastic volatility and other affine jump diffusion processes. Oper Res 54(2):217–231
Lord R, Koekkoek R, Dijk DV (2010) A comparison of biased simulation schemes for stochastic volatility models. Quant Financ 10(2):177–194
Glasserman P, Kim K-K (2011) Gamma expansion of the Heston stochastic volatility model. Financ Stochast 15(2):267–296
Zhu J (2011) A simple and accurate simulation approach to the Heston model. J Deriv 18(4):26–36
Okhrin O, Rockinger M, Schmid M (2022) Simulating the Cox-Ingersoll-Ross and Heston processes: Matching the first four moments. J Computat Financ 26(2):1–52
Pan J (1999) Integrated time-series analysis of spot and option prices. In: AFA 2001 New Orleans Meetings
Andersen TG, Benzoni L, Lund J (2002) An empirical investigation of continuous-time equity return models. J Financ 57(3):1239–1284