Maximum Likelihood Estimation of Latent Markov Models Using Closed-Form Approximations[*]

Yacine Aït-Sahalia[†]
Department of Economics
Princeton University and NBER

Chenxu Li[‡]
Guanghua School of Management
Peking University

Chen Xu Li[§]
School of Business
Renmin University of China

This Version: September 25, 2020

Abstract

This paper proposes and implements an efficient and flexible method to compute maximum likelihood estimators of continuous-time models when part of the state vector is latent. Stochastic volatility and term structure models are typical examples. Existing methods integrate out the latent variables using either simulations as in MCMC, or replace the latent variables by observable proxies. By contrast, our approach relies on closed-form approximations to estimate parameters and simultaneously infer the distribution of filters, i.e., that of the latent states conditioning on observations. Without any particular assumption on the filtered distribution, we approximate in closed form a coupled iteration system for updating the likelihood function and filters based on the transition density of the state vector. Our procedure has a linear computational cost with respect to the number of observations, as opposed to the exponential cost implied by the high dimensional integral nature of the likelihood function. We establish the theoretical convergence of our method as the frequency of observation increases and conduct Monte Carlo simulations to demonstrate its performance.

Keywords: Markov vector; diffusion; likelihood; latent state variables; integrating out; Markov Chain Monte Carlo.

JEL classification: C32; C58

[*] The research of Chenxu Li was supported by the Guanghua School of Management, the Center for Statistical Science, the High-performance Computing Platform, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Ministry of Education) at Peking University, as well as the National Natural Science Foundation of China (Grant 71671003). Chen Xu Li is grateful for a graduate scholarship and funding support from the Graduate School of Peking University as well as support from the Bendheim Center for Finance at Princeton University and the School of Business at Renmin University of China. All authors contributed equally.

[†] Address: JRR Building, Princeton, NJ 08544, USA. E-mail address: yacine@princeton.edu.

[‡] Address: Guanghua School of Management, Peking University, Beijing, 100871, PR China. E-mail address: cxli@gsm.pku.edu.cn.

[§] Address: School of Business, Renmin University of China, Beijing, 100872, PR China. E-mail address: lichenxu@rmbs.ruc.edu.cn.

1 Introduction

Markov models with latent factors are commonly used in econometrics in general and financial econometrics in particular. Typical examples include stochastic volatility models with unobservable volatility factors (see, e.g., Heston (1993), Bates (1996), and Christoffersen et al. (2009) among others), term structure models for interest rates with unobservable yield factors (see, e.g., Duffie et al. (2000)) and possibly unspanned stochastic volatility factors (see, e.g., Collin-Dufresne and Goldstein (2002), Collin-Dufresne et al. (2009), and Creal and Wu (2015) among others), as well a large variety of asset pricing factor models used to quantify expected returns. When only part of the state vector is observed, the econometric objective is primarily to estimate the parameters of the model, and possibly to filter the values of the latent states, with the filtering aspect sometimes a necessary step in order to estimate the model. To estimate the model parameters by maximum likelihood, one needs to be able to efficiently evaluate the marginal likelihood function, which is the probability density of the observable variables.

There are two main types of econometric approaches to accomplish this goal. The simplest approach consists in supplementing the model with observable proxies for the latent factors, followed by conducting inference on the augmented model made of the observed state variables and the proxies. Examples of proxies in a stochastic volatility setting would be option prices or implied volatilities; in a term structure setting, bond yields. The proxies are not merely observable variables considered to be good replacements for the latent factors, to be simply plugged in their place. They are typically functions of the latent factors as well as of the parameters of the model, and the likelihood function of the augmented state vector can be properly derived from that of the original state vector, including the additional parameter dependence induced by the transformation from latent factor to proxy: see Aït-Sahalia and Kimmel (2007) and Kanaya and Kristensen (2016) for such an approach to estimate stochastic volatility models and Aït-Sahalia and Kimmel (2010) for term structure models. A limitation of these methods is the need to first come up with suitable proxies, and then derive within the model the parameter-dependent relationship between the proxies and the latent factors. This can be reasonably straightforward when the specification of the model is simple, such as in affine term structure models since these models produce bond yields that are affine functions of the factors, but can be challenging otherwise.

A second and in principle more general approach consists in integrating out the latent factors in order to construct the marginal likelihood function of the observed variables only. A naive method would numerically integrate out the latent variables from the joint probability density of both observable and latent variables, resulting in a high dimensional integral, with dimension equal to the number of observations. Such a task quickly becomes numerically prohibitive for a practical sample with thousands or more observations, even before one attempts to maximize this marginal likelihood over the parameter space, a step which itself requires multiple recomputations of the marginal likelihood along the way. So brute-force numerical integration, while theoretically straightforward, is out of the question as it involves unfeasibly high computational complexity of exponential growth with respect to the number of latent variables to integrate out, i.e., the number of observations. Similarly, filtering consists in estimating the distribution of the latent variable conditioned on the estimated model parameters and the past observations up to the current time. Most non-Gaussian latent Markov models do not admit closed-form filtered distributions. The characterization of filtered distributions turns out again to involve an infinite dimensional problem, unless approximate (but in most cases theoretically unjustified) assumptions are directly imposed on such a distribution, such as being Normal or exponential.[1]

As a result of the impossibility of a direct evaluation, various approximations methods to more effectively build the marginal likelihood and/or filter the latent states have been proposed in the literature (see, e.g., Chen (2003) for a survey). Simulation-based methods include the Expectation-Maximization (EM) algorithm and its various extensions (see, e.g., Dempster et al. (1977) and Little and Rubin (2019)), indirect inference (Gouriéroux et al. (1993) and Smith (1993)), simulated method of moments (Duffie and Singleton (1993) and Gallant and Tauchen (1996)), simulated maximum likelihood methods (Danielsson and Richard (1993), Sandmann and Koopman (1998), Durham (2006), and Kleppe et al. (2014)), Markov Chain Monte Carlo (MCMC thereafter) (see, e.g., Jacquier et al. (1994), Kim et al. (1999), Eraker (2001), Chib et al. (2002), and Johannes and Polson (2010)), and particle filtering algorithms for filtering and sequential parameter learning (see, e.g., Kotecha and Djuric (2003) and Johannes et al. (2009)), among others. Analytical methods include Kalman filtering (see Kalman (1960) for the Kalman filter, Schmidt (1966) for an extended filter, Alspach and Sorenson (1972) for a Gaussian sum filter approximation approach, Julier and Uhlmann (2004) for an unscented filter, and Kawakatsu (2007) for a numerical integration based Gaussian sum filter approximation approach), generalized method of moments with closed-form moment formulae (Melino and Turnbull (1990) and Andersen and Sørensen (1996)), methods based on numerical integration and a piecewise linear approximation of the filtered and one-step predicted filtered densities under state-space models (a class of discrete-time Markov models with a self-driven latent process) in Kitagawa (1987) and Fridman and Harris (1998), approximate likelihood estimation methods based on the explicit characteristic function of transition density under continuous-time affine models (Bates (2006)), closed-form moment-based estimation (Dufour and Valéry (2009)), quasi-maximum likelihood method by employing a multivariate normal density matching the first and second order moments of the transition density or the density of innovation terms (Harvey et al. (1994), Ruiz (1994), and Hurn et al. (2013)).

In this paper, we propose an alternative approximation of the marginal likelihood and the latent

[1] The filtered distribution admits a finite dimensional characterization only in some particular examples; see, e.g., Kalman (1960) and Kalman and Bucy (1961) for the Kalman and Kalman-Bucy filters, Beneš (1981) and Beneš (1985) for Beneš filters, among others.

variables filtered density that applies to general continuous-time models with data that is discretely sampled. We start from the Bayesian filtering formulation of the likelihood and filtered densities. This allows for recursive updates of the likelihood and filtered density as new observations become available, but is still subject to the same two difficulties – the high dimensional integration problem for the evaluation of marginal likelihood function and the infinite dimensional characterization problem for the filtered distributions. We circumvent these difficulties by proposing a method in which the likelihood function and a set of generalized filtered moments are alternately and recursively updated in an efficient way. Our construction employs a closed-form approximation of one-step transition quantities (e.g., marginal transition density) derived from the model with respect to basis functions.

This method has the following advantages. First, we impose no particular assumption on the filtered distributions unlike the Gaussian, mixed Gaussian, or exponential assumptions employed in, e.g., Alspach and Sorenson (1972), Bates (2006), Kawakatsu (2007), and Hurn et al. (2013). Second, our approximation admits a linear growth computational complexity with respect to the number of observations, as opposed to the exponential growth implied by the high dimensional integral nature of the marginal likelihood function. This significant reduction of computational cost facilitates the likelihood maximization procedure, in which the likelihood function must be evaluated for hundreds or thousands of times at various parameter values. Third, we show both theoretically and numerically that our approximations of the marginal likelihood function and generalized filtered moments are accurate and converge as more observations become available. Even if subject to approximation errors in each iteration step, we show that the overall resulting errors for both likelihood and filters tend to be stable rather than propagate or explode along the iteration.

We conduct Monte Carlo simulations using the stochastic volatility model of Heston (1993) as an example. Thanks to the closed-form iteration system, the maximum likelihood estimation procedure can be completed within a few minutes for a path with thousands of observations. To identify the loss of efficiency from not observing the latent volatility factor, we compare our marginal maximum likelihood estimators to the corresponding full-information likelihood estimators in which we observe the true realizations of volatilities. (Of course, full-information estimators are only feasible in simulations since the latent stochastic volatility is not observed in practice). As in other settings, full information likelihood estimators outperform partial information ones: see, e.g., Aït-Sahalia and Mykland (2003). Here, we find that for the parameter solely involved in the observable dynamics and the parameter representing the long-run mean of variance, the marginal estimators and the full-information estimators share similar performance. However, full-information estimators significantly outperform the corresponding marginal ones for the parameters characterizing the mean-reverting speed of variance, the volatility of variance, and the leverage effect.

The rest of this paper is organized as follows. Section 2 sets up the model and proposes our solution – a recursive updating system for likelihood and generalized filtered moments. In Section

3, we discuss the choice of basis functions and practical implementation. Section 4 is devoted to the convergence theorem for the approximations of the marginal likelihood, generalized filtered moments, and maximum likelihood estimators. We numerically analyze the accuracy of the results in Section 5. Section 6 provides Monte Carlo simulation results. Section 7 concludes. Finally, a guide for practical implementation is provided in Appendix A, technical assumptions are given in Appendix B, and proofs are in Appendix C.

2 Likelihood evaluation and latent factor filtering

2.1 Model and notation

Consider a multivariate continuous-time diffusion model with state vector (X_t, Y_t), in which X_t is observed but Y_t is latent. For illustration purposes, we assume that both X_t and Y_t are one-dimensional and follow the stochastic differential equations (SDE thereafter)

$$dX_t = \mu_1(X_t, Y_t; \theta)dt + \sigma_{11}(X_t, Y_t; \theta)dW_{1t} + \sigma_{12}(X_t, Y_t; \theta)dW_{2t}, \tag{1a}$$

$$dY_t = \mu_2(X_t, Y_t; \theta)dt + \sigma_{21}(X_t, Y_t; \theta)dW_{1t} + \sigma_{22}(X_t, Y_t; \theta)dW_{2t}, \tag{1b}$$

where W_{1t} and W_{2t} are two independent one-dimensional standard Brownian motions; the functions μ_1, μ_2, σ_{11}, σ_{12}, σ_{21}, and σ_{22} are sufficiently smooth and satisfy growth conditions such that the solution to this SDE exists and is unique; θ is an unknown parameter vector in an open bounded set $\Theta \subset \mathbb{R}^K$. We denote by $\mathcal{X}$ and $\mathcal{Y}$ the state spaces of the processes X_t and Y_t, respectively. Without loss of generality, we further assume that $\sigma_{12}(x, y; \theta) \equiv 0$.[2] When either the observable process X_t or the latent process Y_t, or both of them, are multidimensional, the method we propose can be generalized by adapting notations from scalars to vectors/matrices.

We assume that the process X_t is observed at equidistant discrete dates $\{t = i\Delta | i = 0, 1, 2, \ldots, n\}$, where Δ represents the time interval between two successive observations. The marginal likelihood

[2] If $\sigma_{12}(x, y; \theta) \neq 0$, one can introduce two independent Brownian motions B_{1t} and B_{2t} by

$$dB_{1t} = \frac{\sigma_{11}(X_t, Y_t; \theta)dW_{1t} + \sigma_{12}(X_t, Y_t; \theta)dW_{2t}}{\sqrt{\sigma_{11}(X_t, Y_t; \theta)^2 + \sigma_{12}(X_t, Y_t; \theta)^2}} \text{ and } dB_{2t} = \frac{-\sigma_{12}(X_t, Y_t; \theta)dW_{1t} + \sigma_{11}(X_t, Y_t; \theta)dW_{2t}}{\sqrt{\sigma_{11}(X_t, Y_t; \theta)^2 + \sigma_{12}(X_t, Y_t; \theta)^2}},$$

respectively, and then rewrite the model as

$$dX_t = \mu_1(X_t, Y_t; \theta)dt + \hat{\sigma}_{11}(X_t, Y_t; \theta)dB_{1t}, \text{ and } dY_t = \mu_2(X_t, Y_t; \theta)dt + \hat{\sigma}_{21}(X_t, Y_t; \theta)dB_{1t} + \hat{\sigma}_{22}(X_t, Y_t; \theta)dB_{2t},$$

where

$$\hat{\sigma}_{11}(x, y; \theta) = \sqrt{\sigma_{11}(x, y; \theta)^2 + \sigma_{12}(x, y; \theta)^2}, \; \hat{\sigma}_{21}(x, y; \theta) = \frac{\sigma_{11}(x, y; \theta)\sigma_{21}(x, y; \theta) + \sigma_{12}(x, y; \theta)\sigma_{22}(x, y; \theta)}{\sqrt{\sigma_{11}(x, y; \theta)^2 + \sigma_{12}(x, y; \theta)^2}},$$

and

$$\hat{\sigma}_{22}(x, y; \theta) = \frac{\sigma_{11}(x, y; \theta)\sigma_{22}(x, y; \theta) - \sigma_{12}(x, y; \theta)\sigma_{21}(x, y; \theta)}{\sqrt{\sigma_{11}(x, y; \theta)^2 + \sigma_{12}(x, y; \theta)^2}}.$$

function $\mathcal{L}(\theta)$ is the joint probability density function of $\{X_{i\Delta}\}_{i=1}^{n}$ given the first observation X_0, i.e.,

$$\mathcal{L}(\theta) = p(X_{n\Delta}, X_{(n-1)\Delta}, X_{(n-2)\Delta}, \cdots, X_{\Delta}|X_0;\theta).$$

To construct $\mathcal{L}(\theta)$, one needs to integrate out the latent variables in the joint density of $\{(X_{i\Delta}, Y_{i\Delta})\}_{i=1}^{n}$ given X_0 :

$$\mathcal{L}(\theta) = \int_{\mathcal{Y}^{n+1}} p(X_{n\Delta}, y_{n\Delta}, X_{(n-1)\Delta}, y_{(n-1)\Delta}, \cdots, X_{\Delta}, y_{\Delta}, y_0|X_0;\theta)dy_{n\Delta}\cdots dy_{\Delta}dy_0,$$

where $\mathcal{Y}$ is the state space of the latent process Y_t. Denote by $p_{(X,Y)}(\Delta, x_{i\Delta}, y_{i\Delta}|x_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)$ the transition density of (X_t, Y_t). Conditioning and applying the Markov property of (X_t, Y_t), we further express $\mathcal{L}(\theta)$ as

$$\begin{aligned}\mathcal{L}(\theta) &= \int_{\mathcal{Y}^{n+1}} p(X_{n\Delta}, y_{n\Delta}, X_{(n-1)\Delta}, y_{(n-1)\Delta}, \cdots, X_{\Delta}, y_{\Delta}|X_0, y_0;\theta)p(y_0|X_0)dy_{n\Delta}\cdots dy_{\Delta}dy_0 \\ &= \int_{\mathcal{Y}^{n+1}} \prod_{i=1}^{n} p_{(X,Y)}(\Delta, X_{i\Delta}, y_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)p(y_0|X_0)dy_{n\Delta}\cdots dy_{\Delta}dy_0, \qquad (2)\end{aligned}$$

which is an integral of dimension $n+1$.

Assuming that all the parameters are identified by the marginal likelihood (identification being model-specific see, e.g., Newey and Steigerwald (1997) for a related setting), the maximum likelihood estimator of θ obtained by maximizing the marginal likelihood function $\mathcal{L}(\theta)$, or equivalently, the marginal log-likelihood function $\ell(\theta) = \log\mathcal{L}(\theta)$, is the marginal maximum likelihood estimator (MMLE, thereafter):

$$\hat{\theta}^{(n,\Delta)}_{\text{MMLE}} = \underset{\theta}{\operatorname{argmax}}\,\mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmax}}\,\ell(\theta).$$

2.2 Bayes updating system

Filtering consists in predicting the distribution of the latent variables $Y_{i\Delta}$ based on the up-to-date available data $\mathbf{X}_{i\Delta} = (X_{i\Delta}, X_{(i-1)\Delta}, \cdots, X_0)$ for any $i \geq 0$ as well as the parameter estimates. We begin by setting up a recursive Bayes updating system for simultaneously updating the conditional likelihood $p(X_{i\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)$ and the filtered density $p(y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)$ (i.e., the density of the random variable $Y_{i\Delta}$ given $\mathbf{X}_{i\Delta}$).

First, by iterative conditioning, the marginal likelihood function $\mathcal{L}(\theta)$ admits the following product form:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n}\mathcal{L}_i(\theta), \text{ with } \mathcal{L}_i(\theta) = p(X_{i\Delta}|\mathbf{X}_{(i-1)\Delta};\theta). \qquad (3)$$

The likelihood update conditional density $\mathcal{L}_i(\theta)$ characterizes the change of the likelihood function $\mathcal{L}(\theta)$ when a new observation $X_{i\Delta}$ becomes available. The calculation of each likelihood update $\mathcal{L}_i(\theta)$ depends on the filtered density $p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)$ according to

$$\mathcal{L}_i(\theta) = \int_{\mathcal{Y}} p_X(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}, \qquad (4)$$

which follows from further conditioning with respect to $Y_{(i-1)\Delta}$ and the Markov property of (X_t, Y_t). The marginal transition density $p_X(\Delta, x|x_0, y_0; \theta)$ is defined by

$$p_X(\Delta, x|x_0, y_0; \theta) = \int_{\mathcal{Y}} p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta)dy, \tag{5}$$

which integrates out the forward latent variable y in the transition density $p_{(X,Y)}$.

Second, it follows from the definition of conditional density that

$$p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta}; \theta) = \frac{p(y_{(i-1)\Delta}, X_{(i-1)\Delta}|\mathbf{X}_{(i-2)\Delta}; \theta)}{p(X_{(i-1)\Delta}|\mathbf{X}_{(i-2)\Delta}; \theta)}.$$

By further conditioning with respect to $Y_{(i-2)\Delta}$ and the definition of $\mathcal{L}_i(\theta)$ in (3), we obtain

$$\begin{aligned}&p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta}; \theta)\\ &= \frac{1}{\mathcal{L}_{i-1}(\theta)} \int_{\mathcal{Y}} p(y_{(i-1)\Delta}, X_{(i-1)\Delta}|\mathbf{X}_{(i-2)\Delta}, y_{(i-2)\Delta}; \theta) p(y_{(i-2)\Delta}|\mathbf{X}_{(i-2)\Delta}; \theta) dy_{(i-2)\Delta}.\end{aligned}$$

Thanks to the Markov property of the model (1a)–(1b), we have

$$p(y_{(i-1)\Delta}, X_{(i-1)\Delta}|\mathbf{X}_{(i-2)\Delta}, y_{(i-2)\Delta}; \theta) = p_{(X,Y)}(\Delta, X_{(i-1)\Delta}, y_{(i-1)\Delta}|X_{(i-2)\Delta}, y_{(i-2)\Delta}; \theta)$$

by dropping all the past observations up to time $(i-2)\Delta$. As a result, we obtain the following equation for updating the filtered densities from $p(y_{(i-2)\Delta}|\mathbf{X}_{(i-2)\Delta}; \theta)$ to $p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta}; \theta)$:

$$\begin{aligned}&p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta}; \theta)\\ &= \frac{1}{\mathcal{L}_{i-1}(\theta)} \int_{\mathcal{Y}} p_{(X,Y)}(\Delta, X_{(i-1)\Delta}, y_{(i-1)\Delta}|X_{(i-2)\Delta}, y_{(i-2)\Delta}; \theta) p(y_{(i-2)\Delta}|\mathbf{X}_{(i-2)\Delta}; \theta) dy_{(i-2)\Delta}.\end{aligned} \tag{6}$$

Now, treating as inputs the initial filter density $p(y_0|X_0; \theta)$,[3] the transition density $p_{(X,Y)}$, and the marginal transition density p_X defined in (5) based on the transition density $p_{(X,Y)}$, the relations (4) and (6) constitute a coupled system for recursively updating the likelihood and filtered density.[4] The transition density of the model steers the whole updating system.

A straightforward idea for implementing the updating system (4) and (6) would be recursive numerical integration. The multiple integral in the original definition (2) is now seemingly split into single integrals in (4) and (6). However, the high dimensional nature of the integral for calculating likelihood remains unchanged – the dimension continues to equal the sample size n. Any brute-force numerical integration algorithm under such a setting has an exponential growth of complexity with respect to n, and is thus impractical to implement. Indeed, for $i = 1, 2, \ldots, n$, suppose each numerical integration in either (4) or (6) requires j times of evaluation of the corresponding integrand at

[3] The initial filter density $p(y_0|X_0; \theta)$ needs to be specified, as part of the initial condition of model (1a)–(1b). If Y_t is stationary, it can for example be set to the stationary marginal density.

[4] Unlike the Bayes updating system (13)–(15) developed for affine models in the Fourier space by Bates (2006), our system (4)–(6) operates in the probability space.

different grid points. Then, to calculate the likelihood update $\mathcal{L}_n(\theta)$ according to (4), it is necessary to compute j times the values of filtered density $p(y_{(n-1)\Delta}|\mathbf{X}_{(n-1)\Delta};\theta)$ at j different grid points $y_{(n-1)\Delta}$. It follows from (6) that, at each grid point $y_{(n-1)\Delta}$, the filtered density $p(y_{(n-1)\Delta}|\mathbf{X}_{(n-1)\Delta};\theta)$ further depends on j values of the filtered density $p(y_{(n-2)\Delta}|\mathbf{X}_{(n-2)\Delta};\theta)$ at j different grid points $y_{(n-2)\Delta}$. Tracing back to the beginning of the recursive calculation, we have to calculate j^n times the initial filtered density $p(y_0|X_0;\theta)$. To avoid such a prohibitive task, we propose and implement in the next section an approximation system, which reduces the computational complexity to linear growth with respect to the sample size n.

2.3 A recursive updating system for likelihood and filter approximations

We begin by approximating the integral for the likelihood update $\mathcal{L}_i(\theta)$ in (4). Assume the marginal transition density $p_X(\Delta,x|x_0,y_0;\theta)$ admits an approximation with respect to the backward latent variable y_0 of the form:

$$p_X^{(J)}(\Delta,x|x_0,y_0;\theta)=\sum_{k=0}^{J}\alpha_k(\Delta,x_0,x;\theta)b_k(y_0;\theta), \tag{7}$$

for some integer order $J\geq 0$, where $\{b_k\}_{k=0}^J$ represent a collection of basis functions and $\{\alpha_k\}_{k=0}^J$ represent the corresponding coefficients. The actual choice of these basis functions, the calculation of the coefficients in closed-form, and the validation of approximation (7) will be discussed below in Section 3.

Plugging approximation (7) into (4), we obtain the following approximation to the likelihood update $\mathcal{L}_i^{(J)}$:

$$\mathcal{L}_i^{(J)}(\theta)=\int_{\mathcal{Y}}p_X^{(J)}(\Delta,x|x_0,y_0;\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}.$$

This expression simplifies to

$$\mathcal{L}_i^{(J)}(\theta)=\sum_{k=0}^{J}\alpha_k(\Delta,X_{(i-1)\Delta},X_{i\Delta};\theta)\mathcal{M}_{k,(i-1)\Delta}(\theta), \tag{8}$$

where $\mathcal{M}_{k,u\Delta}$ is a generalized moment of the conditional distribution of $Y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta}$ defined by

$$\mathcal{M}_{k,u\Delta}(\theta)=\int_{\mathcal{Y}}b_k(y_{u\Delta};\theta)p(y_{u\Delta}|\mathbf{X}_{u\Delta};\theta)dy_{u\Delta},\text{ for }u=0,1,2,\ldots,n. \tag{9}$$

We call $\mathcal{M}_{k,u\Delta}$ a generalized filtered moment.

To compute these generalized filtered moments $\mathcal{M}_{k,(i-1)\Delta}$, we first plug the filter updating equation (6) into (9)

$$\begin{aligned}&\mathcal{M}_{k,(i-1)\Delta}(\theta)\\&=\frac{1}{\mathcal{L}_{i-1}(\theta)}\int_{\mathcal{Y}}p(y_{(i-2)\Delta}|\mathbf{X}_{(i-2)\Delta};\theta)B_k\left(\Delta,X_{(i-1)\Delta}|X_{(i-2)\Delta},y_{(i-2)\Delta},\theta\right)dy_{(i-2)\Delta},\end{aligned} \tag{10}$$

where B_k is defined by

$$B_k(\Delta, x|x_0, y_0; \theta) = \int_{\mathcal{Y}} b_k(y; \theta) p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) dy, \tag{11}$$

for any $k \leq J$. We call B_k the generalized marginal transition moment. It coincides with the marginal transition density p_X, if $b_k(y; \theta) \equiv 1$. Similar to $p_X^{(J)}$ in (7) for approximating the marginal transition density p_X, we assume that the generalized marginal transition moment B_k admits the following Jth order approximation on the same collection of basis functions $\{b_k\}_{k=0}^J$, i.e.,

$$B_k^{(J)}(\Delta, x|x_0, y_0; \theta) = \sum_{j=0}^{J} \beta_{k,j}(\Delta, x_0, x; \theta) b_j(y_0; \theta). \tag{12}$$

The validity of this approximation hinges on the choice of basis functions $\{b_k\}_{k=0}^J$ and will be discussed also in Section 3. Plugging the assumed approximation (12) into (10) in the place of B_k, we obtain the following approximation of the generalized filtered moment $\mathcal{M}_{k,(i-1)\Delta}$:

$$\tilde{\mathcal{M}}_{k,(i-1)\Delta}^{(J)}(\theta) = \frac{1}{\mathcal{L}_{i-1}(\theta)} \sum_{j=0}^{J} \beta_{k,j}^{(J)}(\Delta, X_{(i-2)\Delta}, X_{(i-1)\Delta}; \theta) \mathcal{M}_{j,(i-2)\Delta}(\theta). \tag{13}$$

Finally, to construct a closed recursive system, we replace both of the likelihood update $\mathcal{L}_{i-1}$ and the generalized filtered moments $\mathcal{M}_{j,(i-2)\Delta}$ on the right hand sides of (8) and (13) by their corresponding Jth order approximations, and thus we obtain the following result.

Theorem 1. *For any integer order $J \geq 0$, we have the following recursive updating system for computing the Jth order approximations of the likelihood update $\hat{\mathcal{L}}_i^{(J)}$ and the generalized filtered moments $\hat{\mathcal{M}}_{k,(i-1)\Delta}^{(J)}$:*

$$\hat{\mathcal{L}}_i^{(J)}(\theta) = \sum_{k=0}^{J} \alpha_k(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta) \hat{\mathcal{M}}_{k,(i-1)\Delta}^{(J)}(\theta), \tag{14}$$

for $i \geq 1$ and

$$\hat{\mathcal{M}}_{k,(i-1)\Delta}^{(J)}(\theta) = \frac{1}{\hat{\mathcal{L}}_{i-1}^{(J)}(\theta)} \sum_{j=0}^{J} \beta_{k,j}(\Delta, X_{(i-2)\Delta}, X_{(i-1)\Delta}; \theta) \hat{\mathcal{M}}_{j,(i-2)\Delta}^{(J)}(\theta), \tag{15}$$

for $i \geq 2$ and $0 \leq k \leq J$. The coefficients α_k (resp. $\beta_{k,j}$) involved in (14) (resp. (15)) are determined by approximation (7) (resp. (12)). As a starting point, the approximation $\hat{\mathcal{M}}_{k,0}^{(J)}$ is chosen as the true initial generalized filtered moment $\mathcal{M}_{k,0}$, i.e.,

$$\hat{\mathcal{M}}_{k,0}^{(J)}(\theta) = \mathcal{M}_{k,0}(\theta) = \int_{\mathcal{Y}} b_k(y_0; \theta) p(y_0|X_0; \theta) dy_0. \tag{16}$$

Here, $p(y_0|X_0; \theta)$ represents the initial filtered density, which is specified as an initial condition together with the model (1a)–(1b).

As an immediate consequence of Theorem 1, the marginal likelihood function $\mathcal{L}(\theta)$ and the marginal log-likelihood function $\ell(\theta) = \log \mathcal{L}(\theta)$ admit the following Jth order approximations

$$\hat{\mathcal{L}}^{(J)}(\theta) = \prod_{i=1}^{n} \hat{\mathcal{L}}_i^{(J)}(\theta) \text{ and } \hat{\ell}^{(J)}(\theta) = \sum_{i=1}^{n} \hat{\ell}_i^{(J)}(\theta), \text{ with } \hat{\ell}_i^{(J)}(\theta) = \log \hat{\mathcal{L}}_i^{(J)}(\theta). \tag{17}$$

Finally, we define the induced approximate marginal maximum likelihood estimator (AMMLE thereafter) by

$$\hat{\theta}_{\text{AMMLE}}^{(n,\Delta,J)} = \underset{\theta}{\operatorname{argmax}}\, \hat{\mathcal{L}}^{(J)}(\theta) = \underset{\theta}{\operatorname{argmax}}\, \hat{\ell}^{(J)}(\theta).$$

The convergence of these approximations for the generalized filtered moments, likelihood function, and MMLE will be discussed in Section 4.

As shown in Theorem 1, our recursive updating system does not involve any numerical integration throughout the whole induction procedure, except for the potential numerical integration in (16) for specifying the initial generalized filtered moments $\mathcal{M}_{k,0}$ in the beginning. The closed-form formulae of the coefficients α_k and $\beta_{k,j}$ are derived once before conducting the induction. Then, the implementation of updating equations (14) and (15) merely depends on simple algebraic calculations. This represents a significant computational advantage for the evaluation of the marginal likelihood and facilitates the maximum likelihood estimation procedure, in which the likelihood function must be evaluated many times while searching for its maximizer.

The characterization of the filtered density $p(y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)$ is an infinite dimensional problem, because the argument $y_{i\Delta}$ takes a continuum of values in the state space $\mathcal{Y}$. This essential challenge results in the exponential computation cost as discussed in Section 2.2. Comparing with the exact updating system (4) and (6), the approximation system (14)–(15) in Theorem 1 employs a finite number (even if one can choose the number to be large) of generalized filtered moments to steer the likelihood evaluation procedure, instead of relying on the filtered density $p(y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)$. This approximation method effectively reduces an infinite dimensional problem to a finite dimensional one, and as a consequence, simplifies the computational complexity with respect to the number of observation from exponential growth to linear growth. Indeed, the approximation system allows for iterating a fixed number of generalized filtered moments over time to approximately characterize the filtered density on a continuum domain at each stage. It turns out that once a new observation becomes available, only J more generalized filtered moments and one more likelihood update ought to be computed based on those from the previous stage. Hence, the computational cost reduces to linear, and thus the iteration becomes implementable in practice. We will demonstrate the computational efficiency through numerical experiments in Section 6. Besides, we will show that, in spite of approximations, all the resulting errors are indeed small in terms of both likelihood and generalized filtered moments evaluations in Sections 4 and 5.

In addition to serving as a building block for the recursive evaluation of the likelihood function, the approximations $\hat{\mathcal{M}}_{k,i\Delta}^{(J)}$ provide a prediction of the distribution of the latent variable $Y_{i\Delta}$ given the historical observations on X up to time $i\Delta$. Different choices of basis functions b_k induce different generalized filtered moments $\mathcal{M}_{k,i\Delta}$, which, from different perspectives, characterize the distribution of $Y_{i\Delta}|\mathbf{X}_{i\Delta}$. For example, a monomial basis function $b_k(y;\theta) = y^k$ results in a typical kth order filtered moment, while an indicator basis function $b_k(y;\theta) = 1_{\{c<y\leq d\}}$ leads to the probability of $Y_{i\Delta}|\mathbf{X}_{i\Delta}$ falling into the interval $(c,d]$. We will discuss in more detail the choice of these basis functions and interpret their corresponding generalized filtered moments in Section 3.

The iteration proposed in Theorem 1 provides a filter that is applicable to non-Gaussian models. In Kalman filtering, since all the filtered distributions are Normal, a full characterization of the filtered distribution at each stage requires only the computation and updating of two filtered moments – mean and variance. For non-Gaussian models, the filtered distributions are no longer normal and are indeed generally unknown. Some existing methods assume for ease of computation the normality of filtered distributions, although this is not theoretically guaranteed: see, e.g., the extended Kalman filtering in Chapter 2 of Javaheri (2015), the Fourier-transform-based filtering method for affine models in Bates (2006), and the quasi-maximum likelihood estimation in Hurn et al. (2013), among others. By contrast, we compute a set of generalized filtered moments $\mathcal{M}_{k,i\Delta}$ to characterize the filtered distributions without relying on assumptions on the filtered distribution.

Before closing this section, we propose (but not implement in this paper due to the length limitation) an immediate extension, which broadens the applicability of Theorem 1. Besides the likelihood update $\mathcal{L}_i$, which is essentially based on the density of the conditional distribution $X_{i\Delta}|\mathbf{X}_{(i-1)\Delta}$, any conditional generalized moment of $X_{i\Delta}|\mathbf{X}_{(i-1)\Delta}$ in the form

$$\mathcal{G}_i(\theta) := \mathbb{E}[g(X_{i\Delta};\theta)|\mathbf{X}_{(i-1)\Delta};\theta] \tag{18}$$

for some generalized function g can be easily computed as a by-product of the system (14)–(15). These conditional generalized moments $\mathcal{G}_i(\theta)$ could be applied for various purposes including but not limited to conditional GMM estimation and hypothesis testing. For example, by letting g be an indicator function, e.g., $g(x;\theta) = 1_{(-\infty,c]}(x)$ for some constant c, one is able to compute the conditional cumulative distribution function

$$\mathbb{E}[1_{(-\infty,c]}(X_{i\Delta})|\mathbf{X}_{(i-1)\Delta};\theta] = \mathbb{P}\left(X_{i\Delta} \leq c|\mathbf{X}_{(i-1)\Delta};\theta\right)$$

and then use it in testing the specification of time-series models; see, e.g., Hong and Li (2005) among others.

The evaluation of the conditional generalized moment $\mathcal{G}_i(\theta)$ proceeds as follows: by analogy to the marginal transition density p_X, consider the generalized marginal moment G by

$$G(\Delta, x_0, y_0;\theta) = \int_{\mathcal{X}} g(x';\theta) p_X\left(\Delta, x'|x_0, y_0;\theta\right) dx'.$$

Accordingly, similar to the marginal transition density approximation (7), we assume that $G(\Delta, x_0, y_0;\theta)$

admits the following Jth order approximation

$$G^{(J)}(x_0, y_0; \theta) = \sum_{k=0}^{J} \gamma_k\,(x_0; \theta)\, b_k(y_0; \theta), \tag{19}$$

on the same collection of basis functions $\{b_k\}_{k=0}^{J}$, where the coefficients γ_k can be calculated in closed form. Then, the conditional generalized moment $\mathcal{G}_i(\theta)$ can be recursively calculated by augmenting the approximation system (14)–(15) to the one in the following corollary.

Corollary 1. *For any integer $i \geq 1$ and the same integer order J as in Theorem 1, the conditional generalized moment $\mathcal{G}_i$ defined in (18) can be approximated by*

$$\hat{\mathcal{G}}_i^{(J)}(\theta) = \sum_{k=0}^{J} \gamma_k \left(X_{(i-1)\Delta}; \theta\right) \hat{\mathcal{M}}_{k,(i-1)\Delta}^{(J)}(\theta), \tag{20}$$

where the coefficients γ_k are determined by (19) and the approximate generalized filtered moments $\hat{\mathcal{M}}_{k,(i-1)\Delta}^{(J)}$ are recursively updated according to the approximation system (14)–(15) proposed in Theorem 1.

Proof. See Appendix C.1. □

3 Choice of basis functions and interpretation of generalized filtered moments

The simplest and most natural choice of the basis functions is the family of monomials, i.e., $b_k(y; \theta) = y^k$. In this case, the generalized filtered moment $\mathcal{M}_{k,l\Delta}$ defined by (9) (resp. generalized marginal transition moment B_k defined by (11)) corresponds to the filtered moment (resp. marginal transition moment). Furthermore, the approximation of the marginal transition density p_X in (7) and the approximation of the generalized marginal transition moment B_k in (12) coincide with their Taylor expansions with respect to the latent backward variable y_0 around $y_0 = 0$, respectively, if the coefficients α_k and $\beta_{k,j}$ are chosen as

$$\alpha_k(\Delta, x_0, x; \theta) = \frac{1}{k!}\frac{\partial^k p_X}{\partial y_0^k}(\Delta, x|x_0, 0; \theta) \text{ and } \beta_{k,j}(\Delta, x_0, x; \theta) = \frac{1}{j!}\frac{\partial^j B_k}{\partial y_0^j}(\Delta, x|x_0, 0; \theta). \tag{21}$$

However, these Taylor expansions do not converge in the case of most stochastic volatility models. Take approximation (7) for the marginal transition density p_X as an example. Although the true density p_X is implicit, we can still see the divergence problem from the simplest Euler approximation $\hat{p}_X$, which has the same small-time behavior as p_X. The Euler marginal transition density $\hat{p}_X$ is Normal:

$$\hat{p}_X(\Delta, x|x_0, y_0; \theta) = \frac{1}{\sqrt{2\pi\Delta}\sigma_{11}(x_0, y_0; \theta)} \exp\left\{-\frac{(x - x_0 - \mu_1(x_0, y_0; \theta)\Delta)^2}{2\sigma_{11}(x_0, y_0; \theta)^2\Delta}\right\}. \tag{22}$$

Suppose the volatility function σ_{11} admits a convergent Taylor expansion with respect to y_0 around $y_0 = 0$ with $\sigma_{11}(x_0, 0; \theta) = 0$ as the leading term.[5] Nevertheless, since $y = 0$ is an essential singularity of the function $\exp\{-1/y^2\}$, the composite Taylor expansion of $\exp\{-1/\sigma_{11}(x_0, y_0; \theta)^2\}$ with respect to y_0 around $y_0 = 0$ diverges. As a result, the Taylor expansion of $\hat{p}_X$ diverges.

To avoid the problem caused by the monomial basis functions, we choose basis functions in the form of piecewise monomials instead:

$$b_{(y^{(k)},y^{(k+1)}),l}(y;\theta) = y^l 1_{\{y^{(k)}<y\leq y^{(k+1)}\}}, \text{ for } 0 \leq l \leq r \text{ and } 0 \leq k \leq m-1. \tag{23}$$

Here, the integer r represents the highest order of monomials we use and the grid points $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$ form a partition of the latent space $\mathcal{Y}$. We employ different superscripts and/or subscripts to represent the basis functions b, coefficients α and β, generalized filtered moments $\mathcal{M}$, etc., appearing in Section 2.3. Taking the basis function $b_{(y^{(k)},y^{(k+1)}),l}$ for instance, rather than using a single subscript k as in our previous exposition for simplicity, we need three quantities $y^{(k)}$, $y^{(k+1)}$, and l to determine this piecewise monomial basis function. With this slight change of notation, the algorithm established in Theorem 1 remains the same: We recursively update the likelihood and $m(r+1)$ (previously $J+1$) generalized filtered moments.

Using the piecewise monomial basis functions $b_{(y^{(k)},y^{(k+1)}),l}$, the approximations of the marginal transition density (7) becomes

$$p_X^{(r,\mathbf{y})}(\Delta, x|x_0, y_0; \theta) = \sum_{k=0}^{m-1}\sum_{l=0}^{r} \alpha_{(y^{(k)},y^{(k+1)}),l}(\Delta, x_0, x; \theta) y_0^l 1_{\{y^{(k)}<y_0\leq y^{(k+1)}\}}, \tag{24}$$

where the coefficient function $\alpha_{(y^{(k)},y^{(k+1)}),l}$ in front of the basis function $y_0^l 1_{\{y^{(k)}<y_0\leq y^{(k+1)}\}}$ will be calculated momentarily. Accordingly, the generalized marginal transition moment is specialized as the truncated marginal transition moment

$$B_{(y^{(k)},y^{(k+1)}),l}(\Delta, x|x_0, y_0; \theta) = \int_{\mathcal{Y}} y^l 1_{\{y^{(k)}<y\leq y^{(k+1)}\}} p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) dy.$$

Its approximation (12) is translated to

$$B^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l}(\Delta, x|x_0, y_0; \theta) = \sum_{j=0}^{m-1}\sum_{\ell=0}^{r} \beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}(\Delta, x_0, x; \theta) y_0^\ell 1_{\{y^{(j)}<y_0\leq y^{(j+1)}\}}, \tag{25}$$

where the subscript (resp. superscript) of the coefficient function $\beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}$ corresponds to the index of the truncated marginal transition moment on the left hand side of (25) (resp. the basis function on the right hand side of (25)). Moreover, the definition (9) for the generalized filtered moment is specialized as the following truncated filtered moment

$$\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,u\Delta}(\theta) = \int_{\mathcal{Y}} y_{u\Delta}^l 1_{\{y^{(k)}<y_{u\Delta}\leq y^{(k+1)}\}} p(y_{u\Delta}|\mathbf{X}_{u\Delta}; \theta) dy_{u\Delta}, \tag{26}$$

[5] This is a feature of most stochastic volatility models, for example, those proposed by Stein and Stein (1991), Heston (1993), and Meddahi (2001).

which characterizes the lth order filtered moment on the interval $(y^{(k)}, y^{(k+1)}]$.

We now discuss how to determine the coefficient functions $\alpha_{(y^{(k)},y^{(k+1)}),l}$ and $\beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}$. Their choices can be flexible in general, as long as the approximations (24) and (25) turn out to be more accurate as either $m \to \infty$ (adding more grid points) or $r \to \infty$ (increasing the order of monomials).

In what follows, we propose a method to explicitly choose them by calculating the Lagrange interpolation coefficients of p_X and B_k. Take the marginal transition density p_X for instance. For $y_0 \in (y^{(k)}, y^{(k+1)}]$, (24) is specialized to

$$p_X^{(r,\mathbf{y})}(\Delta, x|x_0, y_0; \theta) = \sum_{l=0}^{r} \alpha_{(y^{(k)},y^{(k+1)}),l}(\Delta, x_0, x; \theta) y_0^l. \tag{27}$$

This can be naturally thought as an rth order polynomial interpolation of $p_X(\Delta, x|x_0, y_0; \theta)$ with respect to y_0 on the interval $y_0 \in (y^{(k)}, y^{(k+1)}]$. To determine the coefficients $\alpha_{(y^{(k)},y^{(k+1)}),l}$, we need the values of function p_X at different $r+1$ subgrid points on the interval $y_0 \in [y^{(k)}, y^{(k+1)}]$.[6] For simplicity, we choose $r+1$ equidistant subgrid points as $(y_0^{(k)}, y_1^{(k)}, \cdots, y_r^{(k)})$ with $y_i^{(k)} = y^{(k)} + (i/r)(y^{(k+1)} - y^{(k)})$ for $i = 0, 1, \ldots, r$. Then, the Lagrange interpolation on the interval $y_0 \in [y^{(k)}, y^{(k+1)}]$ is given by

$$\sum_{i=0}^{r} \left(p_X(\Delta, x|x_0, y_i^{(k)}; \theta) \prod_{0 \leq j \leq r, j \neq i} \frac{y_0 - y_j^{(k)}}{y_i^{(k)} - y_j^{(k)}} \right). \tag{28}$$

The interpolation coefficients $\alpha_{(y^{(k)},y^{(k+1)}),l}$ can be uniquely solved by matching (27) and (28). In case the required input of the marginal transition density $p_X(\Delta, x|x_0, y_i^{(k)}; \theta)$ does not admit a closed-form expression or is difficult to calculate numerically, one may employ an approximation as a compromise, e.g., the Euler approximation $\hat{p}_X$ in (22).

Given the various piecewise polynomial interpolation possibilities for approximating p_X and B_k, we specialize Theorem 1 to the case of piecewise monomial basis functions in the following proposition, and provide a guide for practical implementation in Appendix A.

Proposition 1. *For any integer $r \geq 0$ and grid points $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$, we have the following recursive updating system for computing the $(r, \mathbf{y})$th order approximations of the likelihood update $\hat{\mathcal{L}}_i^{(r,\mathbf{y})}$ and the truncated filtered moments $\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,(i-1)\Delta}$:*

$$\hat{\mathcal{L}}_i^{(r,\mathbf{y})}(\theta) = \sum_{k=0}^{m-1} \sum_{l=0}^{r} \alpha_{(y^{(k)},y^{(k+1)}),l}(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta) \hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,(i-1)\Delta}, \tag{29}$$

for $i \geq 1$ and

$$\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,(i-1)\Delta}(\theta)$$

[6] By the continuity of the marginal transition density p_X and for the purpose of interpolation, we do not distinguish intervals $(y^{(k)}, y^{(k+1)}]$ and $[y^{(k)}, y^{(k+1)}]$.

$$= \frac{1}{\hat{\mathcal{L}}_i^{(r,\mathbf{y})}(\theta)} \sum_{j=0}^{m-1} \sum_{\ell=0}^{r} \beta_{(y^{(k)},y^{(k+1)}),l}^{(y^{(j)},y^{(j+1)}),\ell}(\Delta, X_{(i-2)\Delta}, X_{(i-1)\Delta};\theta)\hat{\mathcal{M}}_{(y^{(j)},y^{(j+1)}),\ell,(i-2)\Delta}^{(r,\mathbf{y})}(\theta), \tag{30}$$

for $i \geq 2$, $0 \leq k \leq m-1$, and $0 \leq l \leq r$. The coefficients $\alpha_{(y^{(k)},y^{(k+1)}),l}$ (resp. $\beta_{(y^{(k)},y^{(k+1)}),l}^{(y^{(j)},y^{(j+1)}),\ell}$) involved in (29) (resp. (30)) are determined at the stage of setting approximation (24) (resp. (25)). As a starting point, the approximation $\hat{\mathcal{M}}_{(y^{(k)},y^{(k+1)}),l,0}^{(r,\mathbf{y})}$ coincides with the true initial truncated filtered moment $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,0}$, i.e.,

$$\hat{\mathcal{M}}_{(y^{(k)},y^{(k+1)}),l,0}^{(r,\mathbf{y})}(\theta) = \mathcal{M}_{(y^{(k)},y^{(k+1)}),l,0}(\theta) = \int_{\mathcal{Y}} y_0^l 1_{\{y^{(k)}<y_0\leq y^{(k+1)}\}} p(y_0|X_0;\theta)dy_0. \tag{31}$$

Here, $p(y_0|X_0;\theta)$ represents the initial filtered density, which is specified as an initial condition together with the model (1a)–(1b).

Note that the iterations (29) and (30) are not based on numerical integration. Indeed, the Lagrange polynomial interpolation is performed only on the transition part $p_X^{(r,\mathbf{y})}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)$ with respect to $y_{(i-1)\Delta}$ but not the whole integrand of

$$\mathcal{L}_i^{(r,\mathbf{y})}(\theta) = \int_{\mathcal{Y}} p_X^{(r,\mathbf{y})}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}.$$

So, our application of the interpolation leads to the truncated filtered moments (26) but does not have the purpose of producing any Newton-Cotes family of quadratures such as Trapezoidal and Simpson rules, which are derived in essence by integrating polynomial interpolants of integrand. Accordingly, the truncated filtered moment (26) is not numerically integrated.

The truncated filtered moments $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ recursively approximated in Proposition 1 have the following interpretation. By definition (26) for the case of $l = 0$, we see that

$$\mathcal{M}_{(y^{(k)},y^{(k+1)}),0,i\Delta}(\theta) = \int_{\mathcal{Y}} 1_{\{y^{(k)}<y_{i\Delta}\leq y^{(k+1)}\}} p(y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)dy_{i\Delta} = \mathbb{P}\left(y^{(k)} < Y_{i\Delta} \leq y^{(k+1)}|\mathbf{X}_{i\Delta};\theta\right).$$

Thus, the zeroth order truncated filtered moment $\mathcal{M}_{(y^{(k)},y^{(k+1)}),0,i\Delta}$ can be interpreted as the probability of the filtered latent variable $Y_{i\Delta}|\mathbf{X}_{i\Delta}$ falling into the interval $(y^{(k)}, y^{(k+1)}]$. Combining all the zeroth order truncated filtered moments for $k = 0, 1, 2 \ldots, m-1$, consider the piecewise cumulative distribution function (CDF thereafter) $F_{i\Delta,\mathbf{y}}$ defined by

$$\begin{aligned} F_{i\Delta,\mathbf{y}}(y;\theta) &= \sum_{k=0}^{m-1} \mathbb{P}\left(y^{(k)} < Y_{i\Delta} \leq y^{(k+1)}|\mathbf{X}_{i\Delta};\theta\right) \frac{\min(y, y^{(k+1)}) - y^{(k)}}{y^{(k+1)} - y^{(k)}} 1_{\{y>y^{(k)}\}} \\ &\equiv \sum_{k=0}^{m-1} \mathcal{M}_{(y^{(k)},y^{(k+1)}),0,i\Delta}(\theta) \frac{\min(y, y^{(k+1)}) - y^{(k)}}{y^{(k+1)} - y^{(k)}} 1_{\{y>y^{(k)}\}}, \end{aligned}$$

and its corresponding approximation $\hat{F}_{i\Delta,\mathbf{y}}^{(r)}(y;\theta)$:

$$\hat{F}_{i\Delta,\mathbf{y}}^{(r)}(y;\theta) = \sum_{k=0}^{m-1} \hat{\mathcal{M}}_{(y^{(k)},y^{(k+1)}),0,i\Delta}^{(r,\mathbf{y})}(\theta) \frac{\min(y, y^{(k+1)}) - y^{(k)}}{y^{(k+1)} - y^{(k)}} 1_{\{y>y^{(k)}\}}. \tag{32}$$

Here, the approximate truncated filtered moments $\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),0,i\Delta}$ can be computed according to Proposition 1.

In addition to the above inference on the filtered distribution, the sum of truncated filtered moments (over k) of the same order l provides an estimate of the lth filtered moment. Indeed, by definition (26), we find that the truncated integral $M_{i\Delta,l,(y^{(0)},y^{(m)})}(\theta)$ given by

$$M_{i\Delta,l,(y^{(0)},y^{(m)})}(\theta) = \int_{\mathcal{Y}} y_{i\Delta}^l 1_{\{y^{(0)}<y_0\leq y^{(m)}\}} p(y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)dy_{i\Delta} = \sum_{k=0}^{m-1} \mathcal{M}_{(y^{(k)},y^{(k+1)}),l,i\Delta}(\theta), \tag{33}$$

converges to the true (but generally unknown) lth filtered moment $M_{i\Delta,l}(\theta)$ given by

$$M_{i\Delta,l}(\theta) = \int_{\mathcal{Y}} y_{i\Delta}^l p(y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)dy_{i\Delta},$$

as the leftmost grid point $y^{(0)}$ (resp. rightmost grid point $y^{(m)}$) approaches to the left (resp. right) boundary of $\mathcal{Y}$. Then, replacing the true truncated filtered moments $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ in (33) by their corresponding approximations $\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$, we use the sum

$$\hat{M}^{(r,\mathbf{y})}_{i\Delta,l,(y^{(0)},y^{(m)})}(\theta) = \sum_{k=0}^{m-1} \hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,i\Delta}(\theta), \tag{34}$$

to approximate the filtered moment $M_{i\Delta,l}(\theta)$.

As discussed above, we employ polynomial interpolation, a choice among other potential methods, to obtain the coefficient function $\alpha_{(y^{(k)},y^{(k+1)}),l}$ and $\beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}$. The main merit of this approach is that it makes the calculation of the interpolated coefficients simple and numerically efficient. In particular, given the closed-form formulae or closed-form approximation formulae of the marginal transition density p_X and the truncated marginal transition moment $B_{(y^{(k)},y^{(k+1)}),l}$, the interpolated coefficients can be computed in closed form. Additionally, from the theory of polynomial interpolation, the approximation error on each interval $(y^{(k)},y^{(k+1)}]$ reduces to zero as the length of the interval shrinks to zero. Given a sufficiently large range of y_0 for piecewise polynomial interpolations, we can shrink the length of each interval by introducing denser grid points $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$, i.e., enlarging m. If both p_X and $B_{(y^{(k)},y^{(k+1)}),l}$ can be exactly computed at each $y_i^{(k)}$ (this is true for Gaussian models), our piecewise interpolated polynomials converge as $m \to \infty$, and as a result, the algorithm converges. Otherwise, when p_X and/or $B_{(y^{(k)},y^{(k+1)}),l}$ cannot be exactly computed, e.g., for most of the non-Gaussian or non-affine continuous-time models, one needs to use various approximations or expansions to compute them before applying polynomial interpolations. Although subject to errors, embedding such approximations or expansions for one-step transition quantities are usually inevitable and unproblematic as an initial step for estimating sophisticated models with latent factors; see, e.g., Johannes et al. (2009), Hurn et al. (2013), and Kleppe et al. (2014) among others. Similar to Johannes et al. (2009), we discretize the model (1a)–(1b) via the Euler scheme and then compute p_X and $B_{(y^{(k)},y^{(k+1)}),l}$ based on their Euler approximations.[7]

The coupled iteration system (29)–(30) in Proposition 1 is nontrivial, even if a straightforward Euler approximation is employed as the initial step to calculate the one-step transition quantities p_X and $B_{(y^{(k)},y^{(k+1)}),l}$. First, as mentioned in the beginning of this section, the Euler approximation (22) of the marginal transition density p_X also confronts the singularity problem at $y_0 = 0$. Thus, as analyzed above, piecewise monomial basis functions and the related technique of piecewise polynomial interpolation become necessary. Second, it is not generally the case that the filtered variables $Y_{i\Delta}|\mathbf{X}_{i\Delta}$ is simply Normal, even if one treats the Euler-discretized model as the true model; the distribution $Y_{i\Delta}|\mathbf{X}_{i\Delta}$ is still unknown. Hence, we need to resort to the whole procedure developed in this section, as opposed to assuming a known filtered distribution such as normal or exponential, to approximate the likelihood and filters.

4 Convergence of the approximations of likelihood, filters, and MMLE

Under the general basis functions $\{b_k\}_{k=0}^{J}$ and technical assumptions provided in Appendix B, we establish the following result on the convergence of our approximations for the generalized filtered moments $\mathcal{M}_{k,i\Delta}$, likelihood update $\mathcal{L}_i$, log-likelihood update ℓ_i, log-likelihood function ℓ, as well as the resulting AMMLE. The case of piecewise monomial basis functions, which we employ in practice, can be viewed as a special case. For clarity, we explicitly emphasize the dependence on Δ for the generalized filtered moments $\mathcal{M}_{k,i\Delta}(\Delta,\theta)$, likelihood update $\mathcal{L}_i(\Delta,\theta)$, log-likelihood update $\ell_i(\Delta,\theta)$, and log-likelihood $\ell(\Delta,\theta)$, as well as their approximations throughout this section.

Theorem 2. *Under Assumptions 1–10, for any fixed number of observations n, we have the following convergence:*

$$\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \xrightarrow{p} 0,\ \sup_{\theta\in\Theta}\left|\frac{\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)-\mathcal{L}_i(\Delta,\theta)}{\mathcal{L}_i(\Delta,\theta)}\right| \xrightarrow{p} 0, \tag{35}$$

$$\sup_{\theta\in\Theta}\left|\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right| \xrightarrow{p} 0,\ \textit{and}\ \sup_{\theta\in\Theta}\left|\hat{\ell}^{(J)}(\Delta,\theta)-\ell(\Delta,\theta)\right| \xrightarrow{p} 0, \tag{36}$$

as $\Delta \to 0$ and $J \to \infty$, simultaneously. Here, the column vectors $\mathcal{M}_{i\Delta}(\Delta,\theta)$ and $\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)$ collect

[7] There exists a vast of literature discussing the closed-form approximations/expansions of the transition density, which accordingly induce approximations/expansions of p_X and $B_{(y^{(k)},y^{(k+1)}),l}$ after integrating the forward latent variable y. However, these density approximations, to the best of our knowledge, become inevitably inaccurate when the latent backward variable y_0 is close to the singularity. As a consequence, the induced approximations for p_X and $B_{(y^{(k)},y^{(k+1)}),l}$ turn out to be unreliable.

all the generalized filtered moments and their approximations according to

$$\begin{aligned}\mathcal{M}_{i\Delta}(\Delta,\theta) &= (\mathcal{M}_{0,i\Delta}(\Delta,\theta),\mathcal{M}_{1,i\Delta}(\Delta,\theta),\cdots,\mathcal{M}_{J,i\Delta}(\Delta,\theta))^{\intercal},\\ \hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta) &= (\hat{\mathcal{M}}_{0,i\Delta}^{(J)}(\Delta,\theta),\hat{\mathcal{M}}_{1,i\Delta}^{(J)}(\Delta,\theta),\cdots,\hat{\mathcal{M}}_{J,i\Delta}^{(J)}(\Delta,\theta))^{\intercal};\end{aligned}$$

the notation $\|\cdot\|_1$ denotes the L_1-norm for column vectors, and $\xrightarrow{p}$ represents the convergence in probability. Denote by $\hat{\theta}_{MMLE}^{(n,\Delta)}$ and $\hat{\theta}_{AMMLE}^{(n,\Delta,J)}$ the corresponding MMLE and AMMLE obtained by maximizing $\ell(\Delta,\theta)$ and $\hat{\ell}^{(J)}(\Delta,\theta)$, respectively. Then, we have

$$\hat{\theta}_{AMMLE}^{(n,\Delta,J)}-\hat{\theta}_{MMLE}^{(n,\Delta)}\xrightarrow{p}0, \tag{37}$$

as $\Delta\to 0$ and $J\to\infty$, simultaneously. Furthermore, as $n\to\infty$, there exist sequences $\Delta_n\to 0$ and $J_n\to\infty$, such that

$$\sup_{\theta\in\Theta}\left|\hat{\ell}^{(J_n)}(\Delta_n,\theta)-\ell(\Delta_n,\theta)\right|\xrightarrow{p}0 \text{ and } \hat{\theta}_{AMMLE}^{(n,\Delta_n,J_n)}-\hat{\theta}_{MMLE}^{(n,\Delta_n)}\xrightarrow{p}0. \tag{38}$$

Proof. See Appendix C.2. □

Intuitively, the approximation error has two sources: the Euler discretization error from approximating the marginal transition density p_X and the generalized marginal transition moment B_k, and the basis approximation error (7) and (12). The former error decreases to zero as the time interval Δ shrinks to 0, while, for any fixed Δ, the latter error decreases to zero as the number of basis functions $J+1$ increases to ∞. The interpretation of the limit $J\to\infty$ varies. When the piecewise monomial basis functions introduced in Section 3 are employed, if the state space $\mathcal{Y}$ of the latent process is bounded, $J\to\infty$ translates into a denser set of grid points $\mathbf{y}=(y^{(0)},y^{(1)},\cdots,y^{(m)})$, where $y^{(0)}$ (resp. $y^{(m)}$) is set as the lower (resp. upper) bound of $\mathcal{Y}$. Otherwise, if $\mathcal{Y}$ is unbounded, $J\to\infty$ means a simultaneously denser and wider set of grid points.

Theorem 2 establishes the convergence of AMMLE to MMLE in the sense of (37) or (38). No conclusion is drawn for the convergence from MMLE to the true values of parameters, since this is not the objective of the paper. If MMLE converges to the true value for fixed n as $\Delta\to 0$ (resp. $n\to\infty$ and $\Delta_n\to 0$), one can further claim the convergence of AMMLE to the true value in the same sense as (37) (resp. (38)).

5 Numerical accuracy

In this section, we verify the accuracy of the approximations in terms of likelihood evaluation and latent factor filtering in the context of two examples: the stochastic volatility model of Heston (1993) in Section 5.1 and a bivariate Ornstein-Uhlenbeck model with stochastic drift in Section 5.2.

5.1 The Heston model

Assume that the observations are generated by the stochastic volatility model of Heston (1993):

$$dX_t = \left(\mu - \frac{1}{2}Y_t\right) dt + \sqrt{Y_t}dW_{1t}, \tag{39a}$$

$$dY_t = \kappa(\alpha - Y_t)dt + \xi\sqrt{Y_t}[\rho dW_{1t} + \sqrt{1-\rho^2}dW_{2t}], \tag{39b}$$

where W_{1t} and W_{2t} are independent standard Brownian motions. Here, the positive parameters κ, α, and ξ describe the speed of mean-reversion, the long-run mean, and the volatility of the latent process Y_t, respectively. We assume that Feller's condition holds: $2\kappa\alpha \geq \xi^2$. The parameter $\rho \in [-1, 1]$ measures the instantaneous correlation between innovations in X_t and Y_t.Denote by $\theta = (\mu, \kappa, \alpha, \xi, \rho)^\intercal$ the collection of all model parameters.

We simulate a time series of (X_t, Y_t) consisting of $n = 2,500$ consecutive values at the daily frequency, i.e., with time increment $\Delta = 1/252$, by subsampling higher frequency data generated by the Euler scheme. We initialize each path at $X_0 = \log 100$ and sample the first latent variable Y_0 from the stationary distribution of Y_t – the Gamma distribution with shape parameter $\omega = 2\kappa\alpha/\xi^2$ and scale parameter $\delta = \xi^2/(2\kappa)$. Accordingly, we choose the initial filtered density $p(y_0|X_0;\theta)$ as the Gamma density, i.e.,

$$p(y_0|X_0;\theta) = \frac{\delta^{-\omega}}{\Gamma(\omega)} y_0^{\omega-1} e^{-y_0/\delta} 1_{\{y_0 \geq 0\}}. \tag{40}$$

Here, Γ represents the Gamma function, $\Gamma(\omega) = \int_0^\infty t^{\omega-1}e^{-t}dt$. The parameter values are $\mu = 0.05$, $\kappa = 3$, $\alpha = 0.1$, $\xi = 0.25$, and $\rho = -0.7$. For likelihood evaluation and latent factor filtering, we use only observations on X_t.

Since the marginal transition density approximation (7) is subject to the aforementioned singularity problem, we resort to the piecewise monomial basis functions proposed in Section 3. For illustration, the highest order of piecewise monomials r is set to be 3. So throughout the iteration (29)–(30), only truncated filtered moments of orders 0 (piecewise CDF), 1 (piecewise mean), 2, and 3 will be recursively computed. Given the grid points $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$, we approximate the marginal transition density p_X and truncated marginal transition moments $B_{(y^{(k)},y^{(k+1)}),l}$ based on the Euler discretization of model (39a)–(39b), and then apply piecewise cubic polynomial interpolations (24) and (25) with $r = 3$ to obtain the coefficients $\alpha_{(y^{(k)},y^{(k+1)}),l}$ and $\beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}$. As a starting point of the recursive computation, the initial truncated filtered moments $\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,0}$ in (31) are calculated according to the initial filtered density (40), i.e.,

$$\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,0}(\theta) = \frac{\delta^l}{\Gamma(\omega)}\bar{\Gamma}\left(1 + \omega, \frac{y^{(k)}}{\delta}, \frac{y^{(k+1)}}{\delta}\right).$$

Here, $\bar{\Gamma}$ represents the incomplete Gamma function defined by $\bar{\Gamma}(a, z_1, z_2) = \int_{z_1}^{z_2} t^{a-1}e^{-t}dt$.

Even if the Heston model (39a)–(39b), as an affine process, is relatively analytically tractable, it is difficult to implement benchmarks for filter-related statistics, e.g., the truncated filtered moments $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ and the likelihood update $\mathcal{L}_i$ among others. Take $\mathcal{L}_i$ as an example. Based on the discussion at the end of Section 2.2, the computational complexity via the true iteration system (4) and (6) exponentially grows with respect to i. Even worse, to obtain the exact values of the transition density $p_{(X,Y)}$, which enters into the integrands in (4) and (6), additional efforts of numerical integrations would be required, as $p_{(X,Y)}$ is explicit only up to the Fourier transform inversion. The absence of benchmarks prevents us from comparing our approximations with numerically exact values. Alternatively, in what follows, we examine the convergence of the approximations by comparing the change of values in two successive orders.

The only exception lies in that the zeroth order filtered moment $M_{i\Delta,0}$, which is identical to 1 for any $i \geq 0$, can serve as a benchmark for checking convergence. As its approximation, the cumulated truncated filtered moment $\hat{M}^{(r,\mathbf{y})}_{i\Delta,0,(y^{(0)},y^{(m)})}$ given in (34) ought to be close to 1 at any stage i. For now, we choose a sufficiently wide range of the grid points $(y^{(0)}, y^{(m)}]$ as a safeguard, such that for any $i \geq 0$ and $y_{i\Delta} \notin (y^{(0)}, y^{(m)}]$, the value of filtered density $p(y_{i\Delta}|X_{i\Delta};\theta)$ is close to zero and thus can be ignored.[8] Referring to the stationary density (40) with mean $\alpha = 0.1$ and standard deviation $\xi\sqrt{\alpha/(2\kappa)} = 0.032$, we choose $(y^{(0)}, y^{(m)}]$ as $(0.01, 0.3]$. For any positive integer m, we simply set $\mathbf{y}$ as equidistant grids in $(0.01, 0.3]$ while holding the leftmost grid point $y^{(0)}$ (resp. rightmost grid point $y^{(m)}$) at 0.01 (resp. 0.3). More precisely, we have $y^{(k)} = 0.01 + 0.29k/m$, for $k = 0, 1, 2, \ldots, m$. With equidistant choice of grid points within $(0.01, 0.3]$, we write

$$\hat{M}^{(r,m)}_{i\Delta,l}(\theta) = \hat{M}^{(r,\mathbf{y})}_{i\Delta,l,(y^{(0)},y^{(m)})},\ \hat{F}^{(r)}_{i\Delta,m}(y;\theta) = \hat{F}^{(r)}_{i\Delta,\mathbf{y}}(y;\theta),\ \hat{\mathcal{L}}^{(r,m)}_i(\theta) = \hat{\mathcal{L}}^{(r,\mathbf{y})}_i(\theta), \tag{41a}$$

and

$$\hat{\ell}^{(r,m)}_i(\theta) = \log \hat{\mathcal{L}}^{(r,m)}_i(\theta),\ \hat{\ell}^{(r,m)}(\theta) = \sum_{i=1}^{n} \hat{\ell}^{(r,m)}_i(\theta), \tag{41b}$$

for any $i \geq 0$, $m \geq 1$, and $0 \leq l \leq r$.

Figure 1 plots the convergence of the approximation $\hat{M}^{(3,m)}_{i\Delta,0}$ at various stages i as m increases from 10 to 50 and leads to the following implications. First, the approximation errors are always below the level of 0.01, for all $i = 1, 2, \ldots, n$ and $m = 10, 20, \ldots, 50$. Even the lowest order approximation $\hat{M}^{(3,10)}_{i\Delta,0}$ provides reasonable estimates at all times. Second, for each order of approximation m, the error stably propagates and does not increase or explode as more observations become available. Third, for any fixed date $i\Delta$, although the range $(0.01, 0.3]$ is fixed, the approximation error tends to decrease as the number of grid points m increases. This suggests that the range $(0.01, 0.3]$ is wide enough.

[8] Intuitively, this stability of the filtered densities hinges on the stationarity and strong ergodicity of the latent process Y_t. Then, independent of i, the filtered densities take tiny values for extreme values of $y_{i\Delta}$. For the Heston model (39a)–(39b), the latent process Y_t is a CIR progress, which is stationary and strongly ergodic under the imposed Feller condition.

Next, we plot the piecewise filtered CDF $\hat{F}_{i\Delta,m}^{(r)}(y;\theta)$ according to (32) in Figures 2–3 at various stages i and with different orders of approximations m. Each panel in Figures 2–3 examines the convergence of the piecewise CDF approximation. Consider the upper panel of Figure 2 as an example. First, for any m, the approximate CDF is monotonically increasing with the left (resp. right) tail approaching to 0 (resp. 1). Second, the approximate CDF converges as m increases.

Similarly to the case of zeroth order filtered moment $M_{i\Delta,0}$, we check the convergence of the approximations for the first, second, and third order filtered moments in Figures 4–6, respectively.[9] Due to the lack of a benchmark as discussed above, we alternatively compare the changes of approximation values in two successive orders, e.g., the change from order $m = 10$ to 20 and that from order $m = 20$ to 30, etc. All these figures show the convergence and numerical stability of the approximations. Take Figure 4 for the filtered mean (the first order filtered moment) as an example. First, for any stage i, the absolute change of approximation values with two successive orders m decreases as m increases, immediately suggesting the convergence of our approximation. Second, for any order m, the absolute change of approximation values varies around a given level, instead of tending to explode, as i increases.

Likewise, we check the convergence of the approximations for the likelihood update, log-likelihood update, and log-likelihood function. Figure 7 (resp. 8) compares the relative change (resp. absolute change) of the approximation values with two successive orders for the likelihood update (resp. log-likelihood update). We focus on the convergence and numerical stability of these two approximations. For the cases where $m = 10$, 20, 30, 40, and 50, the log-likelihood approximations $\hat{\ell}^{(3,m)}$ are computed as 6331.9685, 6331.9972, 6332.0016, 6332.0033, and 6332.0038, respectively, suggesting convergence.

Finally, we compare the approximation of filtered means with the value of corresponding true latent variables generated together with the observations $\{X_{i\Delta}\}_{i=0}^{n}$ by simulation. According to Theorem 2, the approximation $\hat{M}_{i\Delta,1}^{(3,m)}$ should serve as a reasonable estimate of the true filtered mean $M_{i\Delta,1}$ as $\Delta \to 0$ and $m \to \infty$ simultaneously. On the other hand, by the nature of stochastic volatility models, the latent states can be exactly recovered based on the observations of X_t, as the sampling frequency $1/\Delta$ tends to infinity (see, e.g., Chapter 8 of Aït-Sahalia and Jacod (2014)). Consequently, one expects the approximate filtered mean $\hat{M}_{i\Delta,1}^{(3,m)}$ to be close to the value of true latent state when Δ is sufficiently small. We set the frequency as daily, i.e., $\Delta = 1/252$, and exhibit the comparisons in Figure 9. The upper, middle, and lower panels compare the true states of $Y_{i\Delta}$ (in red) with the approximate filtered means $\hat{M}_{i\Delta,1}^{(3,m)}$ (in black) for $m = 10$, 30, and 50, respectively. In each panel, we additionally provide the confidence intervals, which are constructed by shifting the filtered mean upward and downward twice filtered standard deviation. Here, the filtered standard deviation

[9] We check the convergence of filtered moments $M_{i\Delta,l}$ instead of that of their truncated versions $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$, because the truncated moments $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ with different orders m are not comparable. Indeed, by definition (26), the value of $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ depends on the grid points $y^{(k)}$ and $y^{(k+1)}$, which change with respect to m according to $y^{(k)} = 0.01 + 0.29k/m$.

is given by the square root of the filtered variance, which is approximated by $\hat{M}_{i\Delta,2}^{(3,m)} - (\hat{M}_{i\Delta,1}^{(3,m)})^2$. We find that the approximate filtered mean closely tracks the true states and the difference between them is within the confidence interval. Moreover, we rarely identify significant differences of the filtered means or confidence intervals among these three panels. This suggests that the recovery of latent states can be performed successfully by approximations of filtered means even with a small number of piecewise monomial basis functions.

5.2 A bivariate Ornstein-Uhlenbeck model

We now consider the following bivariate Ornstein-Uhlenbeck (BOU thereafter) model in the form

$$\begin{aligned} dX_t &= \kappa_1(Y_t - X_t)dt + \sigma_1 dW_{1t}, \\ dY_t &= \kappa_2(\alpha - Y_t)dt + \sigma_2 dW_{2t}, \end{aligned}$$

where W_{1t} and W_{2t} are independent standard Brownian motions. Here, the positive parameters κ_1 and σ_1 (resp. κ_2 and σ_2) denote the speed of mean-reversion and the volatility of the observable process X_t (resp. latent process Y_t), respectively. The parameter α characterizes the long-run mean of the latent process Y_t. Denote by $\theta = (\kappa_1, \kappa_2, \alpha, \sigma_1, \sigma_2)^\intercal$ the collection of all model parameters and θ_0 the corresponding true values, which are set as $(1, 3, 0, 0.1, 0.1)^\top$ in the experiments. As a Gaussian vector autoregression model, its likelihood update and filtered distribution/moments are available in closed form and can serve as benchmarks for assessing the accuracy of our approximations.[10]

As in Section 5.1, the general piecewise monomial basis functions $b_{(y^{(k)},y^{(k+1)}),l}(y;\theta)$ in (23) can be used for the BOU model. However, unlike the Heston case, the Euler approximation of the marginal transition density (22) under the BOU model, i.e.,

$$\hat{p}_X(\Delta, x|x_0, y_0; \theta) = \frac{1}{\sqrt{2\pi\Delta}\sigma_1} \exp\left\{-\frac{(x - x_0 - \kappa_1(y_0 - x_0)\Delta)^2}{2\sigma_1^2\Delta}\right\}$$

does not have a singularity at $y_0 = 0$, and we can simply choose the basis functions as monomials instead of piecewise monomials. That is, as proposed at the beginning of Section 3, set $b_k(y;\theta) = y^k$ for $k = 0, 1, 2, \ldots, J$ with some integer order $J \geq 0$. Then, for the marginal transition density p_X and the marginal transition moment B_k, we can employ their Taylor expansions with respect to y_0 to construct the approximations (7) and (12), respectively; the coefficients α_k and $\beta_{k,j}$ are given in (21). Accordingly, the generalized filtered moment $\mathcal{M}_{k,l\Delta}$ (resp. generalized marginal transition moment B_k) degenerates to the typical filtered moment $\mathcal{M}_{k,l\Delta}(\theta) = \int_{\mathcal{Y}} y_{l\Delta}^k p(y_{l\Delta}|\mathbf{X}_{l\Delta};\theta)dy_{l\Delta}$ (resp. marginal transition moment $B_k(\Delta, x|x_0, y_0;\theta) = \int_{\mathcal{Y}} y^k p_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta)dy$).

Figures 10–15 compare the approximations of the likelihood update $\hat{\mathcal{L}}_i(\theta_0)$, log-likelihood update $\hat{\ell}_i(\theta_0)$, and the first four filtered moments $\hat{\mathcal{M}}_{1,i\Delta}$, $\hat{\mathcal{M}}_{2,i\Delta}$, $\hat{\mathcal{M}}_{3,i\Delta}$, $\hat{\mathcal{M}}_{4,i\Delta}$ with the corresponding closed-

[10] Under this model, the Kalman filter applies; see Kalman and Bucy (1961).

form benchmarks.[11] We consider approximations with four different orders, specifically $J = 4$, 6, 8, and 10. We find that as the order increases, the approximation error uniformly decreases at any ordinal of observation i.

Although the BOU model is a useful example for verifying the accuracy of the approximations, we do not include its MMLE, due to its poor empirical performance. Indeed, the MMLE for this model admit large standard deviations unless very large sample sizes (of the order of hundreds of thousands of observations) are available, which is empirically impractical. This phenomenon is due to the specific structure of the model, namely, the fact that the latent process is a stochastic drift rather than a stochastic volatility of the observable process. Even in ideal settings, drift parameters are harder to estimate than volatility ones in moderate to high frequency data. Here, the problem is compounded by the fact that we are attempting to estimate the drift of a (latent) drift. Unless we amplify greatly the speed at which the processes mean-revert, and/or lower the volatility with which they do so, pinning down accurately the level to which they mean-revert is very difficult as a practical matter, despite the apparent simplicity of the BOU model. This difficulty is present already in this model for the true MMLE based on the exact log-likelihood function: it is not a function of the fact that we approximate the MMLE. The estimation performance deteriorates further for any approximate MMLE methods.

6 Monte Carlo simulations

In this section, we conduct Monte Carlo simulations to validate the accuracy of the AMMLE as a potential estimator of θ. We compare our estimators with the full-information ones obtained if we were additionally observing the latent variables $\{Y_{i\Delta}\}_{i=0}^{n}$, so as to identify the loss of efficiency due to the latency of Y_t. (The full-information estimation is of course only feasible in simulated data.) We finally compare our method with the alternative MCMC method in terms of the trade-off between estimation accuracy and computational efficiency.

We use as data generating process the same Heston model as in Section 5 with true parameters $\theta_0 = (\mu_0, \kappa_0, \alpha_0, \xi_0, \rho_0)^\intercal = (0.05, 3, 0.1, 0.25, -0.7)^\top$. We consider two sample frequencies, daily and weekly, corresponding to $\Delta = 1/252$ and $\Delta = 1/52$, respectively, as well as five sample sizes with $n = 1,000,\ 2,500,\ 5,000,\ 10,000$, and $20,000$ observations, respectively. For each sample frequency

[11] For the zeroth order filtered moment $\hat{\mathcal{M}}_{0,i\Delta}$, the approximation $\hat{\mathcal{M}}_{0,i\Delta}^{(J)}$ is identically equal to 1 given the choice of monomial basis functions. Indeed, it follows from (14) and (15) that

$$\hat{\mathcal{M}}_{0,i\Delta}^{(J)}(\theta) = \frac{\sum_{j=0}^{J} \beta_{0,j}(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta)\hat{\mathcal{M}}_{j,(i-1)\Delta}^{(J)}(\theta)}{\sum_{j=0}^{J} \alpha_j(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta)\hat{\mathcal{M}}_{j,(i-1)\Delta}^{(J)}(\theta)}.$$

Here, the coefficients $\beta_{0,j}$ and α_j are the Taylor expansion coefficients of the zeroth order marginal transition moment B_0 and the marginal transition density p_X. Since, by definition, B_0 and p_X are identical to each other given the choice of the monomial basis functions, so are $\beta_{0,j}$ and α_j. It turns out that $\hat{\mathcal{M}}_{0,i\Delta}^{(J)}(\theta) \equiv 1$.

Δ and sample size n, we perform 500 simulation trials and compute 500 AMMLEs.

In each simulation trial, we generate the time series $\{(X_{i\Delta}, Y_{i\Delta})\}_{i=0}^{n}$ in the same way as in Section 5 and calculate the AMMLE by optimizing the approximate marginal log-likelihood function based on the partial observations $\{X_{i\Delta}\}_{i=0}^{n}$. To calculate the approximate log-likelihood function, we employ the piecewise monomial basis functions described in Section 3. Similar to the experiments in Section 5, we set the highest order of piecewise monomials r as 3. Throughout the optimization procedure, the objective function (approximate log-likelihood) needs to be evaluated at various parameter values. Instead of using fixed grid points, we design an algorithm to adaptively choose grid points $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$, depending on different values of parameter θ. For example, if the long-run mean α of Y_t increases, the grid points $\mathbf{y}$ ought to move rightward. Details for the algorithm are in Appendix A.2. Besides computing the AMMLE, for each sample, we estimate the asymptotic variance matrix $V(\theta_0)$ as the inverse of the marginal Fisher information:

$$\hat{V}(\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}}) = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{\ell}_i^{(3,m)}(\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}})}{\partial \theta} \frac{\partial \hat{\ell}_i^{(3,m)}(\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}})}{\partial \theta^{\intercal}} \right)^{-1}. \tag{42}$$

The asymptotic standard deviations of the AMMLE can be consistently estimated by the square root of the diagonal entries of $\hat{V}(\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}})/n$.

For comparison purposes, in each simulation trial, we additionally compute the full-information maximum likelihood estimator (FMLE thereafter) by using the complete sample $\{(X_{i\Delta}, Y_{i\Delta})\}_{i=0}^{n}$:

$$\hat{\theta}^{(n,\Delta)}_{\text{FMLE}} = \underset{\theta}{\operatorname{argmax}}\, \ell_f(\theta), \text{ with } \ell_f(\theta) = \log \mathcal{L}_f(\theta),$$

where $\mathcal{L}_f$ represents the joint density of $\{(X_{i\Delta}, Y_{i\Delta})\}_{i=1}^{n}$ conditioning on (X_0, Y_0), i.e.,

$$\mathcal{L}_f(\theta) = \prod_{i=1}^{n} p_{(X,Y)}(\Delta, X_{i\Delta}, Y_{i\Delta} | X_{(i-1)\Delta}, Y_{(i-1)\Delta}; \theta).$$

The loss of efficiency between the MMLE relative to the FMLE measures the loss of information due to the latent nature of Y_t. Since the full-information log-likelihood function ℓ_f is generally implicit, various approximations/expansions have been proposed based on the approximation of (log) transition density (see, e.g., Aït-Sahalia (2002), Aït-Sahalia (2008) and Li (2013) among others), resulting in an approximate FMLE (AFMLE thereafter). In this section, we use the log-likelihood approximation and the induced AFMLE proposed in Aït-Sahalia (2008) with a log transition density expansion up to the second order of Δ, i.e., the expansion $\tilde{l}^{(2)}_{(X,Y)}(x, y|x_0, y_0, \Delta; \theta)$ introduced in Aït-Sahalia (2008). Both the log-likelihood approximation and the AFMLE have been validated to be accurate for small Δ. Similar to the marginal case, we estimate the asymptotic variance matrix $V_f(\theta_0)$ for AFMLE using the inverse of Fisher information:

$$\hat{V}_f(\hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}) = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \tilde{l}^{(2)}_{(X,Y)}(X_{i\Delta}, Y_{i\Delta} | X_{(i-1)\Delta}, Y_{(i-1)\Delta}, \Delta; \hat{\theta}^{(n,\Delta)}_{\text{AFMLE}})}{\partial \theta} \right.$$

$$\times \frac{\partial \tilde{l}^{(2)}_{(X,Y)}(X_{i\Delta}, Y_{i\Delta}|X_{(i-1)\Delta}, Y_{(i-1)\Delta}, \Delta; \hat{\theta}^{(n,\Delta)}_{\text{AFMLE}})}{\partial \theta^{\intercal}} \Bigg)^{-1} . \tag{43}$$

Again, the asymptotic standard deviations of AFMLE can be consistently estimated by the square root of the diagonal entries of $\hat{V}_f(\hat{\theta}^{(n,\Delta)}_{\text{AFMLE}})/n$.

We summarize the Monte Carlo simulation results in Tables 1 and 2, for daily and weekly frequencies, respectively. Take Table 1 as an example. Panels A, B, and C show the results for AMMLE, AFMLE, and their differences respectively. In Panel A (resp. B), for each parameter and each sample size, the bias and finite-sample standard deviation (in parenthesis) is calculated based on 500 AMMLEs (resp. AFMLEs), while the asymptotic standard deviation (in bracket) is calculated as the mean of 500 sample-based standard deviations resulting from (42) (resp. (43)). In Panel C, for each parameter and each sample size, the bias (resp. finite-sample standard deviation in parenthesis) represents the mean (resp. standard deviation) of 500 estimators of $\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}} - \hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}$.

From Tables 1 and 2, we have the following observations. First, from Panel A of Tables 1 and 2, the estimation bias of each parameter is significantly less than the finite-sample standard deviation, for each combined scenario of frequency and sample size. This suggests the accuracy of our MMLE approximation algorithm. Although subject to e.g., Euler approximation and polynomial interpolation errors in the log-likelihood evaluation procedure, the resulting AMMLE for each parameter is not significantly biased away from the corresponding true value of that parameter. Second, under the same choice of frequency, sample size, and parameter, the finite-sample standard deviation of the AMMLE for each parameter is greater than that of the AFMLE, as expected given the loss of information from not observing $\{Y_{i\Delta}\}_{i=0}^{n}$. Third, for the marginal (resp. full-information) estimation in Panel A (resp. B) of Tables 1 and 2, the asymptotic standard deviations calculated from (42) (resp. (43)) are close to the corresponding finite-sample standard deviation, suggesting that the sample-based approximation of the standard deviations is a reasonable estimator of the standard errors.

Next, we plot for each parameter the root mean squared errors of the AMMLE, AFMLE, and their differences in Figure 16 (resp. 17). From the upper left and middle left panels of Figures 16 and 17, we find that for the parameters μ and α, the AMMLE performs slightly worse than the AFMLE. Indeed, additionally observing $\{Y_{i\Delta}\}_{i=0}^{n}$ does not provide significantly more information for inference regarding μ, since it only appears in the observable dimension according to (39a). For the long-run mean parameter α of the latent process Y_t, it can be surprisingly recovered well based solely on the partial observations $\{X_{i\Delta}\}_{i=0}^{n}$. A heuristic reason for this effect is as follows: the dynamics (39a) imply that $\sum_{i=1}^{n}(X_{i\Delta} - X_{(i-1)\Delta})^2$ converges to $\int_0^{n\Delta} Y_t dt$ as $\Delta \to 0$, which indicates that the discrete quadratic variation is a reasonable estimate of the integrated variance in $[0, n\Delta]$. Furthermore, according to the dynamics (39b), the integrated variance can be approximated by $\alpha n\Delta$. By matching the discrete quadratic variation with $\alpha n\Delta$, we solve α as $(\sum_{i=1}^{n}(X_{i\Delta} - X_{(i-1)\Delta})^2)/(n\Delta)$, which is an approximate estimator of α based solely on the sample path of X_t.

On the other hand, from the upper right, middle right, and lowest panels of Figures 16 and 17, we find that for the parameters κ, ξ, and ρ, the AFMLE significantly outperforms the corresponding AMMLE. In other words, observations $\{Y_{i\Delta}\}_{i=0}^{n}$ play an important role for inference on these three parameters. Consider the parameter ρ as an example, which characterizes the instantaneous correlation between the change of X_t and the change of Y_t. Without directly observing the realizations of Y_t, it is not surprisingly more difficult to accurately estimate the correlation parameter ρ.

In addition to the estimation accuracy, we examine the computational efficiency of the approximations for likelihood evaluation and marginal maximum likelihood estimation. Figure 18 plots the average time cost of one-time likelihood evaluation for various sample size n. Figure 18 shows that the time cost almost linearly grows with respect to the sample size. This numerical finding matches the theoretical analysis on the linear computational complexity provided after Theorem 1. Moreover, the likelihood evaluation is fast: even for $20,000$ observations, the approximation of log-likelihood can be completed within 3 CPU seconds on average. In addition to the likelihood evaluation, we plot the average time cost for one-time marginal maximum likelihood estimation in Figure 19. As shown in Figure 19, the average time cost almost linearly grows with respect to the sample size. With $20,000$ observations, the complete estimation procedure can be completed within 12 minutes (720 CPU seconds) on average.

Finally, we compare our method with the MCMC method (see, e.g., Jacquier et al. (1994) and Johannes and Polson (2010)) in terms of their respective estimation accuracy and computational efficiency. We measure the estimation accuracy of each model parameter by the relative root mean squared error (RRMSE), defined as

$$\mathrm{RRMSE}(\vartheta) = \frac{\mathrm{RMSE}(\vartheta)}{|\vartheta|}, \tag{44}$$

for each parameter ϑ in $(\mu, \kappa, \alpha, \xi, \rho)$. As shown in Figure 20, for each parameter, we find that our method performs better in regard to the accuracy/efficiency trade-off, typically resulting in an improvement by a factor of at least 10 in computational time for a given level of desired accuracy.

7 Conclusions and future directions

We propose and implement an efficient and flexible method for maximum likelihood estimation in continuous-time models with latent state variables. Avoiding the exponential-growth computational complexity with respect to the number of observations directly implied by the high dimensional integral nature of marginal likelihood, we propose an efficient method for approximating the likelihood function with a linear-growth complexity. The log-likelihood function is evaluated via a closed-form iteration system without any numerical integration or simulation, and proves to be numerically accurate, efficient, and stable. It is possible to perform a complete maximum likelihood estimation within several minutes on average in samples containing thousands of observations. The iteration system allows one to infer at the same time the filtered distributions of the latent variables without imposing any extraneous assumptions. We establish theoretically the convergence of the approximations for generalized filtered moments, log-likelihood, and the resulting approximate marginal maximum likelihood estimator, and validate these results in simulations.

The method can be extended in a number of directions. First, the coupled iteration system for likelihood evaluation and latent factor filtering can be further augmented for the purpose of calculating any conditional generalized moments of the observable variable, as illustrated at the end of Section 2.3. Second, the method can in principle be extended to cover models with jumps. Third, the convergence theorem in this paper is a starting point for establishing the asymptotic properties of the approximate marginal maximum likelihood estimator under either the in-fill asymptotic scheme ($\Delta \to 0$ and fixed $n\Delta$, thereby excluding unindentified parameters such as drift ones) or a double asymptotic scheme ($\Delta \to 0$ and $n\Delta \to \infty$ simultaneously), to be compared with those of the full information maximum likelihood estimator under the same sampling schemes. Finally, the marginal likelihood estimators we constructed can be employed for specification testing and conditional moment tests by adapting the results of Newey (1985) to the present setting.

References

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

Aït-Sahalia, Y., 2008. Closed-form likelihood expansions for multivariate diffusions. Annals of Statistics 36, 906–937.

Aït-Sahalia, Y., Jacod, J., 2014. High Frequency Financial Econometrics. Princeton University Press.

Aït-Sahalia, Y., Kimmel, R., 2007. Maximum likelihood estimation of stochastic volatility models. Journal of Financial Economics 83, 413–452.

Aït-Sahalia, Y., Kimmel, R., 2010. Estimating affine multifactor term structure models using closed-form likelihood expansions. Journal of Financial Economics 98, 113–144.

Aït-Sahalia, Y., Mykland, P. A., 2003. The effects of random and discrete sampling when estimating continuous-time diffusions. Econometrica 71, 483–549.

Alspach, D. L., Sorenson, H. W., 1972. Nonlinear Bayesian estimation using Gaussian sum approximations. IEEE Transactions on Automatic Control 17 (4), 439–448.

Andersen, T. G., Sørensen, B. E., 1996. GMM estimation of a stochastic volatility model: A Monte Carlo study. Journal of Business & Economic Statistics 14, 328–352.

Bates, D. S., 1996. Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. Review of Financial Studies 9, 69–107.

Bates, D. S., 2006. Maximum likelihood estimation of latent affine processes. Review of Financial Studies 19, 909–965.

Beneš, V., 1981. Exact finite-dimensional filters for certain diffusions with nonlinear drift. Stochastics 5, 65–92.

Beneš, V., 1985. New exact nonlinear filters with large Lie algebras. Systems & Control Letters 5, 217–221.

Cai, N., Li, C., Shi, C., 2013. Closed-form expansions of discretely monitored Asian options in diffusion models. Mathematics of Operations Research 39 (3), 789–822.

Chen, Z., 2003. Bayesian filtering: From Kalman filters to particle filters, and beyond. Statistics 182, 1–69.

Chib, S., Nardari, F., Shephard, N., 2002. Markov Chain Monte Carlo methods for stochastic volatility models. Journal of Econometrics 108, 281–316.

Christoffersen, P., Heston, S., Jacobs, K., 2009. The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well. Management Science 55, 1914–1932.

Collin-Dufresne, P., Goldstein, R. S., 2002. Do bonds span the fixed income markets? Theory and evidence for unspanned stochastic volatility. The Journal of Finance 57, 1685–1730.

Collin-Dufresne, P., Goldstein, R. S., Jones, C. S., 2009. Can the volatility of interest rates be extracted from the cross-section of bond yields? An investigation of unspanned stochastic volatility. Journal of Financial Economics 94, 47–66.

Creal, D. D., Wu, J. C., 2015. Estimation of affine term structure models with spanned or unspanned stochastic volatility. Journal of Econometrics 185, 60–81.

Danielsson, J., Richard, J.-F., 1993. Accelerated Gaussian importance sampler with application to dynamic latent variable models. Journal of Applied Econometrics 8, S153–S173.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B 39 (1), 1–22.

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

Duffie, D., Singleton, K. J., 1993. Simulated moments estimation of Markov models of asset prices. Econometrica 61, 929–952.

Dufour, J.-M., Valéry, P., 2009. Exact and asymptotic tests for possibly non-regular hypotheses on stochastic volatility models. Journal of Econometrics 150, 193–206.

Durham, G. B., 2006. Monte Carlo methods for estimating, smoothing, and filtering one- and two-factor stochastic volatility models. Journal of Econometrics 133, 273–305.

Eraker, B., 2001. MCMC analysis of diffusion models with application to finance. Journal of Business and Economic Statistics 19, 177–191.

Fridman, M., Harris, L., 1998. A maximum likelihood approach for non-Gaussian stochastic volatility models. Journal of Business & Economic Statistics 16, 284–291.

Gallant, A. R., Tauchen, G. T., 1996. Which moments to match? Econometric Theory 12, 657–681.

Gouriéroux, C., Monfort, A., Renault, E., 1993. Indirect inference. Journal of Applied Econometrics 8, S85–S118.

Harvey, A., Ruiz, E., Shephard, N., 1994. Multivariate stochastic variance models. The Review of Economic Studies 61, 247–264.

Heston, S., 1993. A closed-form solution for options with stochastic volatility with applications to bonds and currency options. Review of Financial Studies 6, 327–343.

Hong, Y., Li, H., 2005. Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. Review of Financial Studies 18, 37–84.

Hurn, S., Lindsay, K. A., McClelland, A. J., 2013. A quasi-maximum likelihood method for estimating the parameters of multivariate diffusions. Journal of Econometrics 172, 106–126.

Jacquier, E., Polson, N. G., Rossi, P. E., 1994. Bayesian analysis of stochastic volatility models. Journal of Business and Economic Statistics 14, 429–434.

Javaheri, A., 2015. Inside volatility filtering: Secrets of the skew. John Wiley & Sons.

Johannes, M., Polson, N., 2010. MCMC methods for financial econometrics. In: Aït-Sahalia, Y., Hansen, L. P. (Eds.), Handbook of Financial Econometrics. Vol. 2. North-Holland, Amsterdam, pp. 1–72.

Johannes, M. S., Polson, N. G., Stroud, J. R., 2009. Optimal filtering of jump diffusions: Extracting latent states from asset prices. The Review of Financial Studies 22, 2759–2799.

Julier, S. J., Uhlmann, J. K., 2004. Unscented filtering and nonlinear estimation. Proceedings of the IEEE 92, 401–422.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, Series D 82, 34–45.

Kalman, R. E., Bucy, R. S., 1961. New results in linear filtering and prediction theory. Journal of Basic Engineering, Series D 83, 95–108.

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

Karatzas, I., Shreve, S. E., 1991. Brownian Motion and Stochastic Calculus. Springer-Verlag.

Kawakatsu, H., 2007. Numerical integration-based Gaussian mixture filters for maximum likelihood estimation of asymmetric stochastic volatility models. Econometrics Journal 10 (2), 342–358.

Kim, S., Shephard, N., Chib, S., 1999. Stochastic volatility: Likelihood inference and comparison with ARCH models. Review of Economic Studies 65, 361–393.

Kitagawa, G., 1987. Non-Gaussian state-space modeling of nonstationary time series. Journal of the American Statistical Association 82, 1032–1041.

Kleppe, T. S., Yu, J., Skaug, H. J., 2014. Maximum likelihood estimation of partially observed diffusion models. Journal of Econometrics 180, 73–80.

Kotecha, J. H., Djuric, P. M., 2003. Gaussian sum particle filtering. IEEE Transactions on Signal Processing 51 (10), 2602–2612.

Li, C., 2013. Maximum-likelihood estimation for diffusion processes via closed-form density expansions. Annals of Statistics 41, 1350–1380.

Little, R. J., Rubin, D. B., 2019. Statistical analysis with missing data, 3rd Edition. Wiley Series in Probability and Statistics. John Wiley & Sons.

Meddahi, N., 2001. An eigenfunction approach for volatility modeling. Tech. rep., Université de Montréal.

Melino, A., Turnbull, S. M., 1990. Pricing foreign currency options with stochastic volatility. Journal of Econometrics 45, 239–265.

Newey, W. K., 1985. Maximum likelihood specification testing and conditional moment tests. Econometrica 53, 1047–1070.

Newey, W. K., Steigerwald, D. G., 1997. Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. Econometrica 65, 587–599.

Ruiz, E., 1994. Quasi-maximum likelihood estimation of stochastic volatility models. Journal of Econometrics 63, 289–306.

Sandmann, G., Koopman, S. J., 1998. Estimation of stochastic volatility models via Monte Carlo maximum likelihood. Journal of Econometrics 87, 271–301.

Schmidt, S. F., 1966. Application of state-space methods to navigation problems. Advances in Control Systems 3, 293–340.

Smith, A. A., 1993. Estimating nonlinear time series models using simulated vector autoregressions. Journal of Applied Econometrics 8, S63–S84.

Stein, E. M., Stein, J. C., 1991. Stock price distributions with stochastic volatility: An analytic approach. Review of Financial Studies 4, 727–752.

Varadhan, S. R. S., 1967. On the behavior of the fundamental solution of the heat equation with variable coefficients. Communications in Pure and Applied Mathematics 20, 431–455.

Appendix

Appendix A A guide for implementation

In this appendix, we provide a guide for implementation of our closed-form approximations of the likelihood function and filters.

Appendix A.1 The main approximation algorithm

We begin with a set of grids $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$, which can be chosen according to an adaptive algorithm discussed in Appendix A.2. Then, it is straightforward to construct the piecewise monomial basis functions $b_{(y^{(k)},y^{(k+1)}),l}(y;\theta)$ in (23), i.e., $b_{(y^{(k)},y^{(k+1)}),l}(y;\theta) = y^l 1_{\{y^{(k)}<y\leq y^{(k+1)}\}}$, for $0 \leq l \leq r$ and $0 \leq k \leq m-1$. Then, one can use these basis functions to construct the truncated filtered moments $\mathcal{M}_{(y^{(k)},y^{(k+1)}),l,u\Delta}(\theta)$ given in (26), as well as derive the approximations of the marginal transition density $p_X^{(r,\mathbf{y})}(\Delta, x|x_0, y_0;\theta)$ and truncated marginal transition moment $B_{(y^{(k)},y^{(k+1)}),l}(\Delta, x|x_0, y_0;\theta)$ given in (24) and (25), respectively. Here, the coefficient functions $\alpha_{(y^{(k)},y^{(k+1)}),l}$ in (24) and $\beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}$ in (25) can be obtained in closed form by polynomial interpolations as illustrated in (27) and (28). Finally, the approximations of the likelihood update $\hat{\mathcal{L}}_i^{(r,\mathbf{y})}$ and the truncated filtered moments $\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ can be computed by iterations (29) and (30) in Proposition 1.

Here are the pseudo codes for implementing the above procedure.

```
Algorithm Closed-form approximations of the likelihood function and filters
Input observations $\{X_{i\Delta}\}_{i=0}^n$ and parameter values $\theta$;
  Apply the adaptive algorithm provided in Appendix A.2 to select the grid points $y^{(0)}$, $y^{(1)}$, $\ldots$, $y^{(m)}$;
  Construct the piecewise monomial basis functions $b_{(y^{(k)},y^{(k+1)}),l}(y;\theta)$ in (23);
  Obtain the closed-form formulae of coefficient functions $\alpha_{(y^{(k)},y^{(k+1)}),l}$ and $\beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}$ by
     polynomial interpolations as illustrated in (27) and (28);
  Calculate the initial truncated filtered moment approximation $\hat{\mathcal{M}}_{(y^{(k)},y^{(k+1)}),l,0}$ according to (31),
     where the initial filter density $p(y_0|X_0;\theta)$ is given in (40);
  $i = 1$;
  Repeat
     Compute the likelihood update approximation $\hat{\mathcal{L}}_i^{(r,\mathbf{y})}$ according to (29);
     Compute the truncated filtered moments approximations $\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ according to (30)
        for all $0 \leq l \leq r$ and $0 \leq k \leq m-1$;
     $i \leftarrow i+1$; continue iteration;
     Otherwise if $i = n+1$;
        Iteration stops;
Output $\hat{\mathcal{L}}_i^{(r,\mathbf{y})}$ and $\hat{\mathcal{M}}^{(r,\mathbf{y})}_{(y^{(k)},y^{(k+1)}),l,i\Delta}$ for all $1 \leq i \leq n$, $0 \leq l \leq r$, and $0 \leq k \leq m-1$.
```

Appendix A.2 An adaptive algorithm for choosing the grid points

We now introduce an algorithm for choosing a set of grids $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$ adaptive to the observations $\{X_{i\Delta}\}_{i=0}^{n}$ as well as the values of parameters θ searched in the optimization procedure. For illustrations, we take the Heston model (39a)–(39b) as an example. The algorithm consists of two steps. In the first step, we determine the range of the grid points, i.e., the leftmost grid point $y^{(0)}$ and the rightmost grid point $y^{(m)}$. In the second step, we compute the grid points between $y^{(0)}$ and $y^{(m)}$ via a closed-form iteration.

Step 1 – Determine the range of the grid points. Similar to the method employed in Section 5, we determine the range by referring to the stationary density of the latent process Y_t, i.e., the Gamma density given in (40). Denote by $F_G(\cdot;\omega,\delta)$ the corresponding CDF of this Gamma density, i.e.,

$$F_G(y;\omega,\delta) = \int_0^y \frac{\delta^{-\omega}}{\Gamma(\omega)} z^{\omega-1} e^{-z/\delta} 1_{\{z\geq 0\}} dz, \text{ with } \omega = \frac{2\kappa\alpha}{\xi^2} \text{ and } \delta = \frac{\xi^2}{2\kappa},$$

and $F_G^{-1}(\cdot;\omega,\delta)$ the quantile function, i.e., inverse function of $F_G(\cdot;\omega,\delta)$. To ensure the range of grid points is sufficiently large, we set

$$y^{(0)} = F_G^{-1}(\lambda;\omega,\delta) \text{ and } y^{(m)} = F_G^{-1}(1-\lambda;\omega,\delta),$$

where λ is a quantile parameter close to 0. In Section 6, we let $\lambda = 10^{-7}$. For now, the notation $y^{(m)}$ just represents the rightmost grid point. The value of m will be determined in the next step.

Step 2 – Compute the rest of grid points $y^{(1)}$, $y^{(2)}$, ..., $y^{(m-1)}$ by iterations. The idea for choosing the grid points hinges on the piecewise polynomial interpolation (24) for the Euler approximation of the marginal transition density $\hat{p}_X$, which, under the Heston model, is given by

$$\hat{p}_X(\Delta, x|x_0, y_0;\theta) = \frac{1}{\sqrt{2\pi\Delta y_0}} \exp\left\{-\frac{1}{2y_0\Delta}\left(x - x_0 - \left(\mu - \frac{1}{2}y_0\right)\Delta\right)^2\right\}. \tag{A.1}$$

Now, regarding $y^{(0)}$ as the first grid point, we will discuss how to choose $y^{(1)}$, $y^{(2)}$, ..., $y^{(m)}$ one after another by iterations.

By the theory of Lagrange interpolation, the third order polynomial interpolation with respect to $y' \in [y^{(0)}, y^{(1)}]$ on the interval $[y^{(0)}, y^{(1)}]$ has the following remainder

$$R(\Delta, x|x_0, y';\theta) = \frac{\hat{p}_X^{(4)}(\Delta, x|x_0, \varsigma;\theta)}{4!} \prod_{q=0}^{3} \left(y' - y^{(0)} - \frac{q}{3}\left(y^{(1)} - y^{(0)}\right)\right),$$

for some $\varsigma \in [y^{(0)}, y^{(1)}]$, where $\hat{p}_X^{(4)}$ represents the fourth order derivative of $\hat{p}_X(\Delta, x|x_0, y_0;\theta)$ with respect to y_0 that can be calculated in closed form according to (A.1). By computation, the upper bound of $R(\Delta, x|x_0, y';\theta)$ is given by

$$\left|R(\Delta, x|x_0, y';\theta)\right| \leq \frac{(y^{(1)} - y^{(0)})^4}{4! \times 3^4} \sup_{y' \in [y^{(0)}, y^{(1)}]} |\hat{p}_X^{(4)}(\Delta, x|x_0, y';\theta)|.$$

Given the observations $\{X_{i\Delta}\}_{i=0}^{n}$, to ensure the polynomial interpolation $\hat{p}_X(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y';\theta)$ with respect to y' is accurate enough at any stage i, we determine the grid point $y^{(1)}$ by controlling the above upper bound via

$$\frac{(y^{(1)}-y^{(0)})^4}{4!\times 3^4}\sup_{y'\in[y^{(0)},y^{(1)}],\ i=1,2,\ldots,n}|\hat{p}_X^{(4)}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y';\theta)|\leq\phi,$$

where ϕ is a threshold parameter specified before applying this adaptive algorithm. In Section 6, we let $\phi = 10^{-4}$. Since the length of interval $[y^{(0)}, y^{(1)}]$ is small, by the continuity of $\hat{p}_X^{(4)}$, we approximate the supremum $\sup_{y'\in[y^{(0)},y^{(1)}]}|\hat{p}_X^{(4)}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y';\theta)|$ by $|\hat{p}_X^{(4)}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y^{(0)};\theta)|$ and solve the grid point $y^{(1)}$ from

$$\frac{(y^{(1)}-y^{(0)})^4}{4!\times 3^4}\sup_{i=1,2,\ldots,n}|\hat{p}_X^{(4)}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y^{(0)};\theta)|=\phi.$$

As a consequence, given the first grid point $y^{(0)}$, the next one $y^{(1)}$ is determined in closed form by

$$y^{(1)}=\left(\frac{4!\times 3^4\phi}{\sup\limits_{i=1,2,\ldots,n}|\hat{p}_X^{(4)}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y^{(0)};\theta)|}\right)^{\frac{1}{4}}.$$

Suppose we have chosen the $(k+1)$th grid point $y^{(k)}$ satisfying $y^{(k)} < y^{(m)}$. Then, following a similar discussion, we choose the $(k+2)$th grid point $y^{(k+1)}$ by

$$y^{(k+1)}=\left(\frac{4!\times 3^4\phi}{\sup\limits_{i=1,2,\ldots,n}|\hat{p}_X^{(4)}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y^{(k)};\theta)|}\right)^{\frac{1}{4}}. \tag{A.2}$$

In case $y^{(k+1)}$ exceeds the right edge of the range determined in *Step 1*, i.e., $y^{(m)}$, we let $y^{(k+1)} = y^{(m)}$ and $m = k + 1$ at the same time, and thus the algorithm for choosing grid points stops. Otherwise, we keep $y^{(k+1)}$ as calculated from (A.2), and the iteration (A.2) proceeds for choosing the next grid point.

Here are the pseudo codes for implementing the above algorithm on adaptively choosing the grid points given observations $\{X_{i\Delta}\}_{i=0}^{n}$ and parameter values θ.

Algorithm Adaptive grid points selection

```
**Input** $\{X_{i\Delta}\}_{i=0}^{n}$, $\theta=(\mu,\kappa,\alpha,\xi,\rho)^\intercal$, $\lambda$, and $\phi$;
  $y^{(0)}=F_G^{-1}(\lambda;\omega,\delta)$ with $\omega=2\kappa\alpha/\xi^2$ and $\delta=\xi^2/(2\kappa)$;
  $k=0$;
  **Repeat**
    Compute $y^{(k+1)}$ according to (A.2);
    **If** $y^{(k+1)}<F_G^{-1}(1-\lambda;\omega,\delta)$;
      $k\leftarrow k+1$; continue iteration;
    **Otherwise if** $y^{(k+1)}\geq F_G^{-1}(1-\lambda;\omega,\delta)$;
      $y^{(k+1)}\leftarrow F_G^{-1}(1-\lambda;\omega,\delta)$; $m\leftarrow k+1$; iteration stops;
**Output** $y^{(0)}$, $y^{(1)}$, ..., $y^{(m)}$.
```

Appendix B Technical assumptions and lemmas

Before providing technical assumptions and lemmas, we introduce the following auxiliary notations. To compact notations, we introduce the following vectors and matrices. The collection of basis functions $b^{(J)}$ is defined by

$$b^{(J)}(y;\theta) = (b_0(y;\theta), b_1(y;\theta), \cdots, b_J(y;\theta))^\intercal.$$

Subject to these basis functions, we introduce the following vector-valued functions B and $\hat{B}$ as the collections of generalized marginal transition moments and their approximations, e.g., Euler approximations,

$$\begin{aligned} B(\Delta, x|x_0, y_0;\theta) &= (B_0(\Delta, x|x_0, y_0;\theta), B_1(\Delta, x|x_0, y_0;\theta), \cdots, B_J(\Delta, x|x_0, y_0;\theta))^\intercal, \\ \hat{B}(\Delta, x|x_0, y_0;\theta) &= (\hat{B}_0(\Delta, x|x_0, y_0;\theta), \hat{B}_1(\Delta, x|x_0, y_0;\theta), \cdots, \hat{B}_J(\Delta, x|x_0, y_0;\theta))^\intercal. \end{aligned}$$

Furthermore, we let $\alpha^{(J)}$ be the row vector collecting the approximation coefficients α_k in (7), i.e.,

$$\alpha^{(J)}(\Delta, x, x_0;\theta) = (\alpha_0(\Delta, x, x_0;\theta), \alpha_1(\Delta, x, x_0;\theta), \cdots, \alpha_J(\Delta, x, x_0;\theta)),$$

and $\beta^{(J)}$ be the $(J+1)$-dimensional square matrix collecting the approximation coefficients $\beta_{k,j}$ in (12), i.e.,

$$\beta^{(J)}(\Delta, x, x_0;\theta) = \begin{pmatrix} \beta_{0,0}(\Delta, x, x_0;\theta) & \beta_{0,1}(\Delta, x, x_0;\theta) & \cdots & \beta_{0,J}(\Delta, x, x_0;\theta) \\ \beta_{1,0}(\Delta, x, x_0;\theta) & \beta_{1,1}(\Delta, x, x_0;\theta) & \cdots & \beta_{1,J}(\Delta, x, x_0;\theta) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{J,0}(\Delta, x, x_0;\theta) & \beta_{J,1}(\Delta, x, x_0;\theta) & \cdots & \beta_{J,J}(\Delta, x, x_0;\theta) \end{pmatrix}.$$

Finally, for any column vector $v = (v_1, v_2, \cdots, v_m)^\intercal$, $\|v\|_1$ represents its L_1-norm, i.e., $\|v\|_1 = \sum_{i=1}^m |v_i|$; for any row vector $v = (v_1, v_2, \cdots, v_m)$, its L_1-norm $\|v\|_1$ is given by $\|v\|_1 = \max_{1\leq i\leq m} |v_i|$; for any m-dimensional square matrix $A = (a_{ij})$, its L_1-norm $\|A\|_1$ is given by $\|A\|_1 = \max_{1\leq i\leq m} \sum_{j=1}^m |a_{ij}|$.

We now list our technical assumptions and lemmas. In particular, for assumptions, we provide necessary discussions and/or justifications.

Assumption 1. *The state space $\mathcal{Y}$ is compact with upper bound U, i.e., $|y| \leq U$ for any $y \in \mathcal{Y}$. Moreover, there exists a positive constant $a_0 > 0$, such that for any $y \in \mathcal{Y}$, either $[y, y + a_0] \subset \mathcal{Y}$ or $[y - a_0, y] \subset \mathcal{Y}$.*

Assumption 2. *For each integer $k \geq 1$, the kth order derivatives in (x, y) of the functions $\mu_1(x, y;\theta)$, $\mu_2(x, y;\theta)$, $\sigma_{11}(x, y;\theta)$, $\sigma_{21}(x, y;\theta)$, and $\sigma_{22}(x, y;\theta)$ are uniformly bounded for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where μ_1, μ_1, σ_{11}, σ_{21}, and σ_{22} are the coefficient functions of model (1a)–(1b).*

Assumption 3. *(Uniform ellipticity condition) For any bivariate vector $v = (v_1, v_2)^\intercal \in \mathbb{R}^2$, there exist positive constants a_1 and a_2, such that*

$$\inf_{(x,y,\theta)\in\mathcal{X}\times\mathcal{Y}\times\Theta} v^\intercal \sigma(x,y;\theta)\sigma(x,y;\theta)^\intercal v \geq a_1(v_1^2 + v_2^2), \tag{B.1a}$$

$$\sup_{(x,y,\theta)\in\mathcal{X}\times\mathcal{Y}\times\Theta} v^\intercal \sigma(x,y;\theta)\sigma(x,y;\theta)^\intercal v \leq a_2(v_1^2 + v_2^2), \tag{B.1b}$$

where $\sigma(x, y; \theta)$ is the disperse matrix, i.e.,

$$\sigma(x,y;\theta) = \begin{pmatrix} \sigma_{11}(x,y;\theta) & 0 \\ \sigma_{21}(x,y;\theta) & \sigma_{22}(x,y;\theta) \end{pmatrix}.$$

For now, we assume Assumptions 1–3 hold in order to simplify the proof of Theorem 2. When the compactness of the latent state space $\mathcal{Y}$ (in Assumption 1), the boundedness of function derivatives (in Assumption 2), and/or the uniform ellipticity condition (in Assumption 3) do not hold, a smoothing technique can be applied to the model (1a)–(1b); see, e.g., Appendix E in Cai et al. (2013). The main idea is to construct a smoothed model satisfying Assumptions 1 and 2, and more importantly, the density of the smoothed process converges to that of the original one in probability.

Lemma 1. *Under Assumptions 1 and 3, for any $\Delta > 0$, there exist positive constants M_1, M_2, M_3, α_1, α_2, α_3, and λ, such that for any $\Delta > 0$ and $(x, y, x_0, y_0) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{Y}$, we have*

$$\sup_{\theta\in\Theta} p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) \leq \frac{\Delta^{-1}}{M_1} \exp\left\{-\alpha_1 \frac{(x-x_0)^2 + (y-y_0)^2}{\Delta}\right\}, \tag{B.2}$$

and

$$\begin{aligned} &\inf_{\theta\in\Theta} p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) \\ &\geq \frac{\Delta^{-1}}{M_2} \exp\left\{-\alpha_2 \frac{(x-x_0)^2 + (y-y_0)^2}{\Delta}\right\} - \frac{\Delta^{-1+\lambda}}{M_3} \exp\left\{-\alpha_3 \frac{(x-x_0)^2 + (y-y_0)^2}{\Delta}\right\}. \end{aligned} \tag{B.3}$$

Moreover, for the marginal transition density p_X, there exist positive constants C_1, C_2, and C_3, such that

$$\sup_{\theta\in\Theta} p_X(\Delta, x|x_0, y_0; \theta) \leq C_1 \Delta^{-\frac{1}{2}} \exp\left\{-\frac{\alpha_1 (x-x_0)^2}{\Delta}\right\}, \tag{B.4}$$

and

$$\inf_{\theta\in\Theta} p_X(\Delta, x|x_0, y_0; \theta) \geq C_2 \Delta^{-\frac{1}{2}} \exp\left\{-\frac{\alpha_2 (x-x_0)^2}{\Delta}\right\} - C_3 \Delta^{-\frac{1}{2}+\lambda} \exp\left\{-\frac{\alpha_3 (x-x_0)^2}{\Delta}\right\}. \tag{B.5}$$

Proof. The upper and lower bounds (B.2) and (B.3) follow from a similar discussion as Theorem 2.1 in Varadhan (1967). By the definition of marginal transition density, we have

$$p_X(\Delta, x|x_0, y_0; \theta) \leq \int_{\mathcal{Y}} \sup_{\theta\in\Theta} p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) dy$$

$$
\begin{aligned}
&\leq \frac{\Delta^{-1}}{M_1}\exp\left\{-\frac{\alpha_1(x-x_0)^2}{\Delta}\right\}\int_{\mathcal{Y}}\exp\left\{-\frac{\alpha_1(y-y_0)^2}{\Delta}\right\}dy\\
&\leq \frac{\sqrt{2\pi}}{M_1\sqrt{\alpha_1}}\Delta^{-\frac{1}{2}}\exp\left\{-\frac{\alpha_1(x-x_0)^2}{\Delta}\right\}.
\end{aligned}
$$

Then, (B.4) follows by taking supremum on both sides of the above inequality and then denoting by $C_1 = \sqrt{2\pi}/(M_1\sqrt{\alpha_1})$. On the other hand,

$$
\begin{aligned}
p_X(\Delta, x|x_0, y_0;\theta) &\geq \int_{\mathcal{Y}}\inf_{\theta\in\Theta} p_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta)dy\\
&\geq \frac{\Delta^{-1}}{M_2}\int_{\mathcal{Y}}\exp\left\{-\alpha_2\frac{(x-x_0)^2+(y-y_0)^2}{\Delta}\right\}dy\\
&\quad - \frac{\Delta^{-1+\lambda}}{M_3}\int_{\mathcal{Y}}\exp\left\{-\alpha_3\frac{(x-x_0)^2+(y-y_0)^2}{\Delta}\right\}dy.
\end{aligned}
$$

By Assumption 1, we further deduce

$$
p_X(\Delta, x|x_0, y_0;\theta) \geq \frac{M_4}{2M_2}\Delta^{-\frac{1}{2}}\exp\left\{-\frac{\alpha_2(x-x_0)^2}{\Delta}\right\} - \frac{\sqrt{2\pi}}{M_3\sqrt{\alpha_3}}\Delta^{-\frac{1}{2}+\lambda}\exp\left\{-\frac{\alpha_3(x-x_0)^2}{\Delta}\right\},
$$

where the constant M_4 is given by

$$
M_4 = \int_{|y-y_0|\leq a_0}\frac{1}{\sqrt{\Delta}}\exp\left\{-\frac{\alpha_2(y-y_0)^2}{\Delta}\right\}.
$$

Then, (B.5) follows by taking infimum on both sides of the above inequality and then denoting by $C_2 = M_4/2M_2$ and $C_3 = \sqrt{2\pi}/(M_3\sqrt{\alpha_3})$. □

Lemma 2. *For any $\epsilon > 0$ and integer $i \geq 1$, there exist positive constants $C_{\mathcal{L}}^{\epsilon}$ and $\Delta_{\mathcal{L}}^{\epsilon} > 0$, such that for any $\Delta < \Delta_{\mathcal{L}}^{\epsilon}$, we have*

$$
\mathbb{P}\left(\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta;\theta) \geq C_{\mathcal{L}}^{\epsilon}\Delta^{-\frac{1}{2}}\right) \geq 1-\epsilon. \tag{B.6}
$$

Proof. For any $i \geq 1$, note that $(X_{i\Delta} - X_{(i-1)\Delta})^2 + (Y_{i\Delta} - Y_{(i-1)\Delta})^2 = O_p(\Delta)$. Then, for any $\epsilon > 0$, there exists some $M > 0$, such that

$$
\mathbb{P}\left((X_{i\Delta} - X_{(i-1)\Delta})^2 + (Y_{i\Delta} - Y_{(i-1)\Delta})^2 > M\Delta\right) < \epsilon.
$$

Moreover, we have

$$
\mathbb{P}\left((X_{i\Delta} - X_{(i-1)\Delta})^2 > M\Delta\right) \leq \mathbb{P}\left((X_{i\Delta} - X_{(i-1)\Delta})^2 + (Y_{i\Delta} - Y_{(i-1)\Delta})^2 > M\Delta\right) < \epsilon,
$$

which implies

$$
\mathbb{P}\left((X_{i\Delta} - X_{(i-1)\Delta})^2 \leq M\Delta\right) \geq 1 - \mathbb{P}\left((X_{i\Delta} - X_{(i-1)\Delta})^2 > M\Delta\right) \geq 1-\epsilon.
$$

Now, we prove (B.6). By (B.5) in Lemma 1, we have

$$
\mathcal{L}_i(\Delta;\theta) \geq \int_{\mathcal{Y}}\inf_{\theta\in\Theta} p_X(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)p(y_{(i-1)\Delta}|X_{(i-1)\Delta})dy_{(i-1)\Delta}
$$

$$= C_2\Delta^{-\frac{1}{2}} \exp\left\{-\frac{\alpha_2(X_{i\Delta} - X_{(i-1)\Delta})^2}{\Delta}\right\} - C_3\Delta^{-\frac{1}{2}+\lambda} \exp\left\{-\frac{\alpha_3(X_{i\Delta} - X_{(i-1)\Delta})^2}{\Delta}\right\}.$$

Note that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{L}_i\left(\Delta;\theta\right) \geq C_{\mathcal{L}}^{\epsilon}\Delta^{-\frac{1}{2}}\right)\\
&\geq \mathbb{P}\left(\inf_{\theta\in\Theta} \mathcal{L}_i\left(\Delta;\theta\right) \geq C_{\mathcal{L}}^{\epsilon}\Delta^{-\frac{1}{2}}, (X_{i\Delta} - X_{(i-1)\Delta})^2 \leq M\Delta\right)\\
&= \mathbb{P}\left(\inf_{\theta\in\Theta} \mathcal{L}_i\left(\Delta;\theta\right) \geq C_{\mathcal{L}}^{\epsilon}\Delta^{-\frac{1}{2}}\middle| (X_{i\Delta} - X_{(i-1)\Delta})^2 \leq M\Delta\right) \mathbb{P}\left((X_{i\Delta} - X_{(i-1)\Delta})^2 \leq M\Delta\right)\\
&\geq \mathbb{P}\left(\inf_{\theta\in\Theta} \mathcal{L}_i\left(\Delta;\theta\right) \geq C_{\mathcal{L}}^{\epsilon}\Delta^{-\frac{1}{2}}\middle| (X_{i\Delta} - X_{(i-1)\Delta})^2 \leq M\Delta\right)(1-\epsilon).
\end{aligned}$$

Next, we will choose an appropriate $C_{\mathcal{L}}^{\epsilon}$, such that the last conditional probability is one. Indeed, given $(X_{i\Delta} - X_{(i-1)\Delta})^2 \leq M\Delta$, we have

$$\inf_{\theta\in\Theta} \mathcal{L}_i\left(\Delta;\theta\right) \geq C_2\Delta^{-\frac{1}{2}}e^{-\alpha_2 M} - C_3\Delta^{-\frac{1}{2}+\lambda}.$$

Thanks to the positiveness of the constant λ, there exists $\Delta_{\mathcal{L}}^{\epsilon} > 0$, such that for any $\Delta < \Delta_{\mathcal{L}}^{\epsilon}$, we have

$$C_2\Delta^{-\frac{1}{2}}e^{-\alpha_2 M} - C_3\Delta^{-\frac{1}{2}+\lambda} \geq \frac{1}{2}C_2\Delta^{-\frac{1}{2}}e^{-\alpha_2 M}.$$

Then, (B.6) follows by letting $C_{\mathcal{L}}^{\epsilon} = C_2 e^{-\alpha_2 M}/2$. □

Assumption 4. *For any n observations $\mathbf{X}_{n\Delta}$ with $\Delta > 0$, the L_1-norm of the generalized filtered moments $\left\|\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1$ is uniformly bounded, i.e.,*

$$\sup_{\theta\in\Theta}\sup_{i=1,2,\ldots,n} \left\|\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq C_0, \tag{B.7}$$

for some constant $C_0 > 0$.

If we employ a collection of piecewise monomial basis functions up to the third order in each interval, then Assumption 4 is a straightforward implication of Assumption 1, and thus does not need to be further imposed. Indeed, by employing the piecewise monomial basis functions, we have

$$\begin{aligned}
\left\|\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 &= \sum_{j=0}^{J}\sum_{r=0}^{3}\int_{\mathcal{Y}} \left|y_{i\Delta}\right|^r 1_{(y^{(j)},y^{(j+1)}]}(y_{i\Delta})p(\Delta, y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)dy_{i\Delta}\\
&\leq \sum_{r=0}^{3}\int_{\mathcal{Y}} \left|y_{i\Delta}\right|^r p(\Delta, y_{i\Delta}|\mathbf{X}_{i\Delta};\theta)dy_{i\Delta}.
\end{aligned}$$

By Assumption 1, we have $\left|y_{i\Delta}\right|^r \leq U^r$ and thus

$$\left\|\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq 1 + U + U^2 + U^3.$$

Then, (B.7) follows by letting $C_0 = 1 + U + U^2 + U^3$ and then taking supremum on both sides of the above inequality with respect to θ and i.

Assumption 5. *The marginal transition density p_X and its approximation $\hat{p}_X$ satisfy*

$$\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} |p_X(\Delta,x|x_0,y_0;\theta)-\hat{p}_X(\Delta,x|x_0,y_0;\theta)| \leq \rho_1(\Delta)\Delta^{-\frac{1}{2}}, \tag{B.8}$$

Here, the function $\rho_1(\Delta)$ satisfies $\rho_1(\Delta)\to 0$ as $\Delta\to 0$.

We justify Assumption 5 for the Euler approximation $\hat{p}_X$ given in (22), i.e.,

$$\hat{p}_X(\Delta,x|x_0,y_0;\theta)=\frac{1}{\sqrt{2\pi\Delta}\sigma_{11}(x_0,y_0;\theta)}\exp\left\{-\frac{(x-x_0-\mu_1(x_0,y_0;\theta)\Delta)^2}{2\sigma_{11}(x_0,y_0;\theta)^2\Delta}\right\}. \tag{B.9}$$

To check (B.8), it suffices to show

$$\Delta^{\frac{1}{2}}\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} |p_X(\Delta,x|x_0,y_0;\theta)-\hat{p}_X(\Delta,x|x_0,y_0;\theta)| \to 0, \text{ as } \Delta\to 0. \tag{B.10}$$

According to Assumptions 1–2 and some arguments similar to those for establishing Theorem 2 in Li (2013), the marginal transition density $p_{(X,Y)}$ satisfies

$$\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left|p_X(\Delta,x|x_0,y_0;\theta)-\tilde{p}_X^{(0)}(\Delta,x|x_0,y_0;\theta)\right| = O(1), \tag{B.11}$$

where $\tilde{p}_X^{(0)}$ is the zeroth order expansion of p_X, i.e.,

$$\tilde{p}_X^{(0)}(\Delta,x|x_0,y_0;\theta)=\frac{1}{\sqrt{2\pi\Delta}\sigma_{11}(x_0,y_0;\theta)}\exp\left\{-\frac{(x-x_0)^2}{2\sigma_{11}(x_0,y_0;\theta)^2\Delta}\right\}. \tag{B.12}$$

On the other hand, a comparison of (B.9) and (B.12) implies

$$\begin{aligned}&\hat{p}_X(\Delta,x|x_0,y_0;\theta)-\tilde{p}_X^{(0)}(\Delta,x|x_0,y_0;\theta)\\&=\frac{1}{\sqrt{2\pi\Delta}\sigma_{11}(x_0,y_0;\theta)}\exp\left\{-\frac{(x-x_0)^2}{2\sigma_{11}(x_0,y_0;\theta)^2\Delta}\right\}\left(\exp\left\{\frac{(x-x_0)\mu_1(x_0,y_0;\theta)\Delta}{\sigma_{11}(x_0,y_0;\theta)^2\Delta}\right.\right.\\&\left.\left.-\frac{\mu_1(x_0,y_0;\theta)^2}{2\sigma_{11}(x_0,y_0;\theta)^2}\Delta\right\}-1\right).\end{aligned}$$

If $x\neq x_0$, the first exponent $\exp\{-(x-x_0)^2/(2\sigma_{11}(x_0,y_0;\theta)^2\Delta)\}$ dominates the behavior of the difference $\hat{p}_X-\tilde{p}_X^{(0)}$. As a result, we have $\hat{p}_X-\tilde{p}_X^{(0)}\to 0$ as $\Delta\to 0$. Otherwise, if $x=x_0$, the first exponent $\exp\{-(x-x_0)^2/(2\sigma_{11}(x_0,y_0;\theta)^2\Delta)\}$ is identical to 1, and the second exponent reduces to $\exp\{-\mu_1(x_0,y_0;\theta)^2\Delta/(2\sigma_{11}(x_0,y_0;\theta)^2)\}$, which is approximately $1+O(\Delta)$ when Δ is small. Consequently, the difference $\hat{p}_X-\tilde{p}_X^{(0)}$ is of order $O(\Delta^{1/2})$ and thus converges to zero as $\Delta\to 0$. To sum up, we have

$$\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left|\tilde{p}_X^{(0)}(\Delta,x|x_0,y_0;\theta)-\hat{p}_X(\Delta,x|x_0,y_0;\theta)\right| = o(1), \tag{B.13}$$

Combining (B.11) and (B.13), we obtain

$$\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} |p_X(\Delta,x|x_0,y_0;\theta)-\hat{p}_X(\Delta,x|x_0,y_0;\theta)|$$

$$
\begin{aligned}
&\leq \sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left|p_X(\Delta, x|x_0, y_0;\theta) - \tilde{p}_X^{(0)}(\Delta, x|x_0, y_0;\theta)\right| \\
&+ \sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left|\tilde{p}_X^{(0)}(\Delta, x|x_0, y_0;\theta) - \hat{p}_X(\Delta, x|x_0, y_0;\theta)\right| \\
&= O(1) + o(1) = O(1).
\end{aligned}
$$

Then, (B.10) follows from multiplying $\Delta^{1/2}$ on both sides of the above inequality and then letting $\Delta \to 0$.

Assumption 6. *The marginal transition moments* $B = (B_0, B_1, \cdots, B_J)^\intercal$ *and their approximations* $\hat{B} = (\hat{B}_0, \hat{B}_1, \cdots, \hat{B}_J)^\intercal$ *satisfy*

$$
\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left\|B(\Delta, x|x_0, y_0;\theta) - \hat{B}(\Delta, x|x_0, y_0;\theta)\right\|_1 \leq \rho_2(\Delta)\Delta^{-\frac{1}{2}}, \tag{B.14}
$$

Here, the function $\rho_2(\Delta)$ *satisfies* $\rho_2(\Delta) \to 0$ *as* $\Delta \to 0$.

We justify Assumption 6 when considering piecewise monomial basis functions up to the highest order $r = 3$ and choosing $\hat{B}$ as the Euler approximations. Without loss of generality, we assume $\mathcal{Y} \subset \mathbb{R}_+$. It suffices to show

$$
\Delta^{\frac{1}{2}} \sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left\|B(\Delta, x|x_0, y_0;\theta) - \hat{B}(\Delta, x|x_0, y_0;\theta)\right\|_1 \to 0, \text{ as } \Delta \to 0.
$$

By definition,

$$
\begin{aligned}
&\left\|B(\Delta, x|x_0, y_0;\theta) - \hat{B}(\Delta, x|x_0, y_0;\theta)\right\|_1 \\
&= \sum_{k=0}^{m-1}\sum_{l=0}^{3} \left|B_{(y^{(k)},y^{(k+1)}),l}(\Delta, x|x_0, y_0;\theta) - \hat{B}_{(y^{(k)},y^{(k+1)}),l}(\Delta, x|x_0, y_0;\theta)(\Delta, x|x_0, y_0;\theta)\right| \\
&\leq \sum_{k=0}^{m-1} \int_{\mathcal{Y}} (1 + y + y^2 + y^3) 1_{\{y^{(k)}<y\leq y^{(k+1)}\}} \left|p_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta) - \hat{p}_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta)\right| dy \\
&= \int_{\mathcal{Y}} (1 + y + y^2 + y^3) \left|p_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta) - \hat{p}_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta)\right| dy.
\end{aligned}
$$

Define the set $\mathcal{S}_{x,x_0,y_0,\theta}$ by

$$
\mathcal{S}_{x,x_0,y_0,\theta} = \{y \in \mathcal{Y} : p_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta) \geq \hat{p}_{(X,Y)}(\Delta, x, y|x_0, y_0;\theta)\}.
$$

Then, we have

$$
\begin{aligned}
&\left\|B(\Delta, x|x_0, y_0;\theta) - \hat{B}(\Delta, x|x_0, y_0;\theta)\right\|_1 \\
&\leq B_{\mathcal{S}}(\Delta, x|x_0, y_0;\theta) - \hat{B}_{\mathcal{S}}(\Delta, x|x_0, y_0;\theta) + \hat{B}_{\mathcal{S}^c}(\Delta, x|x_0, y_0;\theta) - B_{\mathcal{S}^c}(\Delta, x|x_0, y_0;\theta),
\end{aligned} \tag{B.15}
$$

where

$$B_{\mathcal{S}}(\Delta, x|x_0, y_0; \theta) = \int_{\mathcal{S}_{x,x_0,y_0,\theta}} (1 + y + y^2 + y^3) p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) dy,$$
$$\hat{B}_{\mathcal{S}}(\Delta, x|x_0, y_0; \theta) = \int_{\mathcal{S}_{x,x_0,y_0,\theta}} (1 + y + y^2 + y^3) \hat{p}_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) dy,$$

and

$$\hat{B}_{\mathcal{S}^c}(\Delta, x|x_0, y_0; \theta) = \int_{\mathcal{Y}\backslash \mathcal{S}_{x,x_0,y_0,\theta}} (1 + y + y^2 + y^3) \hat{p}_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) dy,$$
$$B_{\mathcal{S}^c}(\Delta, x|x_0, y_0; \theta) = \int_{\mathcal{Y}\backslash \mathcal{S}_{x,x_0,y_0,\theta}} (1 + y + y^2 + y^3) p_{(X,Y)}(\Delta, x, y|x_0, y_0; \theta) dy.$$

Similar to the discussions for justifying Assumption 5, we can obtain

$$B_{\mathcal{S}}(\Delta, x|x_0, y_0; \theta) - \hat{B}_{\mathcal{S}}(\Delta, x|x_0, y_0; \theta) = O(1) \text{ and } \hat{B}_{\mathcal{S}^c}(\Delta, x|x_0, y_0; \theta) - B_{\mathcal{S}^c}(\Delta, x|x_0, y_0; \theta) = O(1).$$

Then, (B.14) follows from combining the above two formulae with (B.15).

Assumption 7. *Based on the marginal transition density approximation $\hat{p}_X$, its further approximation satisfies: for any $\Delta > 0$, there exists a $J_{\Delta,1}$, such that for any $J > J_{\Delta,1}$,*

$$\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left| \hat{p}_X(\Delta, x|x_0, y_0; \theta) - \alpha^{(J)}(\Delta, x_0, x; \theta) b^{(J)}(y_0; \theta) \right| \leq \eta_1(\Delta, J_{\Delta,1}) \Delta^{-\frac{1}{2}},$$

Here, the function $\eta_1(\Delta, J_{\Delta,1})$ satisfies $\eta_1(\Delta, J_{\Delta,1}) \to 0$ as $\Delta \to 0$.

Owing to the space limitation, we justify, for simplicity, Assumption 7 and all the rest of assumptions for the case of piecewise constant basis functions, i.e., $r = 0$, with equidistant grid points. In this case, the number of basis function $J + 1$ correspond to the number of piecewise constant basis functions $m + 1$. For the case of $r \geq 1$ or non-equidistant grid points, similar but tedious arguments can be developed.

Under the piecewise constant basis functions, the interpolation coefficient $\alpha_{(y^{(k)},y^{(k+1)}),0}$ is specialized as

$$\alpha_{(y^{(k)},y^{(k+1)}),0}(\Delta, x_0, x; \theta) = \hat{p}_X(\Delta, x|x_0, y^{(k)}; \theta), \tag{B.16}$$

where $\hat{p}_X$ is the Euler approximation given by (B.9). For any interval $(y^{(k)}, y^{(k+1)}]$, the Lagrange interpolation remainder is given by

$$\begin{aligned} &\left| \hat{p}_X(\Delta, x|x_0, y_0; \theta) - \alpha_{(y^{(k)},y^{(k+1)}),0}(\Delta, x_0, x; \theta) \right| \\ &= \left| \frac{\partial \hat{p}_X}{\partial y_0}(\Delta, x|x_0, \varsigma; \theta) \right| (y_0 - y^{(k)}) \\ &\leq \sup_{(x_0,x,\varsigma,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left| \frac{\partial \hat{p}_X}{\partial y_0}(\Delta, x|x_0, \varsigma; \theta) \right| (y^{(k+1)} - y^{(k)}). \end{aligned}$$

According to (B.9), the first order derivative $\partial \hat{p}_X / \partial y_0$ can be calculated explicitly as

$$\begin{aligned}
&\frac{\partial \hat{p}_X}{\partial y_0}(\Delta, x | x_0, \varsigma; \theta) \\
&= \frac{1}{\sqrt{2\pi\Delta}} \exp\left\{ -\frac{(x - x_0 - \mu_1(x_0, \varsigma; \theta)\Delta)^2}{2\sigma_{11}(x_0, \varsigma; \theta)^2 \Delta} \right\} \left[-\frac{\sigma_{11}'(x_0, \varsigma; \theta)}{\sigma_{11}(x_0, \varsigma; \theta)^2} - \frac{1}{2\Delta\sigma_{11}(x_0, \varsigma; \theta)} \right. \\
&\left. \times \left(\frac{2\mu_1'(x_0, \varsigma; \theta)\Delta}{\sigma_{11}(x_0, \varsigma; \theta)^2}(x - x_0 - \mu_1(x_0, \varsigma; \theta)\Delta) - \frac{2\sigma_{11}'(x_0, \varsigma; \theta)}{\sigma_{11}(x_0, \varsigma; \theta)^3}(x - x_0 - \mu_1(x_0, \varsigma; \theta)\Delta)^2 \right) \right],
\end{aligned}$$

where μ_1' and σ_{11}' represent the first order derivatives of μ_1 and σ_{11} with respect to y_0, respectively. By Assumption 3, the function σ_{11} has a positive lower bound $\sqrt{a_1}$. This, together with the boundedness of μ_1' and σ_{11}' imposed in Assumption 2, implies the boundedness of $\partial \hat{p}_X / \partial y_0$, i.e., there exists $C_p > 0$, such that

$$\sup_{(x_0, x, \varsigma, \theta) \in \mathcal{X}^2 \times \mathcal{Y} \times \Theta} \left| \frac{\partial \hat{p}_X}{\partial y_0}(\Delta, x | x_0, \varsigma; \theta) \right| \leq C_p \Delta^{-\frac{1}{2}}.$$

Consequently, we obtain

$$\begin{aligned}
&\sup_{(x_0, x, y_0, \theta) \in \mathcal{X}^2 \times \mathcal{Y} \times \Theta} \left| \hat{p}_X(\Delta, x | x_0, y_0; \theta) - \sum_{k=0}^{m-1} \alpha_{(y^{(k)}, y^{(k+1)}),0}(\Delta, x_0, x; \theta) 1_{\{y^{(k)} < y_0 \leq y^{(k+1)}\}} \right| \\
&\leq C_p \Delta^{-\frac{1}{2}} \max_{0 \leq k \leq m-1} \left(y^{(k+1)} - y^{(k)} \right).
\end{aligned}$$

Since the latent state space $\mathcal{Y}$ is compact and the grid points $\mathbf{y} = (y^{(0)}, y^{(1)}, \cdots, y^{(m)})$ are equidistant, for any $\epsilon > 0$, there exists a sufficiently large integer $\bar{m}_{\Delta,1}^{\epsilon} > 0$, such that for any $m > \bar{m}_{\Delta,1}^{\epsilon}$, we have

$$\max_{0 \leq k \leq m-1} \left(y^{(k+1)} - y^{(k)} \right) < \frac{\epsilon}{C_p},$$

and thus

$$\sup_{(x_0, x, y_0, \theta) \in \mathcal{X}^2 \times \mathcal{Y} \times \Theta} \left| \hat{p}_X(\Delta, x | x_0, y_0; \theta) - \sum_{k=0}^{m-1} \alpha_{(y^{(k)}, y^{(k+1)}),0}(\Delta, x_0, x; \theta) 1_{\{y^{(k)} < y_0 \leq y^{(k+1)}\}} \right| < \epsilon \Delta^{-\frac{1}{2}}.$$

Assumption 8. *Based on the marginal transition moment approximations $\hat{B}$, their further approximations satisfy the following error control. For any $\Delta > 0$, there exists a $J_{\Delta,2}$, such that for any $J > J_{\Delta,2}$,*

$$\sup_{(x_0, x, y_0, \theta) \in \mathcal{X}^2 \times \mathcal{Y} \times \Theta} \left\| \hat{B}(\Delta, x | x_0, y_0; \theta) - \beta^{(J)}(\Delta, x_0, x; \theta) b^{(J)}(y_0; \theta) \right\|_1 \leq \eta_2(\Delta, J_{\Delta,2}) \Delta^{-\frac{1}{2}},$$

Here, the function $\eta_2(\Delta, J_{\Delta,2})$ satisfies $\eta_2(\Delta, J_{\Delta,2}) \to 0$ as $\Delta \to 0$.

Under the piecewise constant basis functions, the interpolation coefficient $\beta_{(y^{(k)}, y^{(k+1)}),0}^{(y^{(j)}, y^{(j+1)}),0}$ is specialized as

$$\beta_{(y^{(k)}, y^{(k+1)}),0}^{(y^{(j)}, y^{(j+1)}),0}(\Delta, x_0, x; \theta) = \hat{B}_{(y^{(k)}, y^{(k+1)}),0}(\Delta, x | x_0, y^{(j)}; \theta). \tag{B.17}$$

It follows that

$$\begin{aligned}
&\left\|\hat{B}(\Delta, x|x_0, y_0; \theta) - \beta^{(J)}(\Delta, x_0, x; \theta) b^{(J)}(y_0; \theta)\right\|_1 \\
&= \sum_{k=0}^{m-1} \left| \hat{B}_{(y^{(k)}, y^{(k+1)}),0}(\Delta, x|x_0, y_0; \theta) - \sum_{j=0}^{m-1} \hat{B}_{(y^{(k)}, y^{(k+1)}),0}(\Delta, x|x_0, y^{(j)}; \theta) 1_{\{y^{(j)} < y_0 \leq y^{(j+1)}\}} \right|.
\end{aligned}$$

The Lagrange interpolation remainder is given by

$$\begin{aligned}
&\left| \hat{B}_{(y^{(k)}, y^{(k+1)}),0}(\Delta, x|x_0, y_0; \theta) - \sum_{j=0}^{m-1} \hat{B}_{(y^{(k)}, y^{(k+1)}),0}(\Delta, x|x_0, y^{(j)}; \theta) 1_{\{y^{(j)} < y_0 \leq y^{(j+1)}\}} \right| \\
&= \left| \sum_{j=0}^{m-1} \frac{\partial \hat{B}_{(y^{(k)}, y^{(k+1)}),0}}{\partial y_0}(\Delta, x|x_0, \varsigma_{k,j}; \theta)(y_0 - y^{(j)}) 1_{\{y^{(j)} < y_0 \leq y^{(j+1)}\}} \right| \\
&\leq \max_{0 \leq j \leq m-1} \left(y^{(j+1)} - y^{(j)} \right) \sum_{j=0}^{m-1} \left| \frac{\partial \hat{B}_{(y^{(k)}, y^{(k+1)}),0}}{\partial y_0}(\Delta, x|x_0, \varsigma_{k,j}; \theta) \right| 1_{\{y^{(j)} < y_0 \leq y^{(j+1)}\}},
\end{aligned}$$

where $\varsigma_{k,j} \in (y^{(j)}, y^{(j+1)}]$. It turns out that

$$\begin{aligned}
&\left\|\hat{B}(\Delta, x|x_0, y_0; \theta) - \beta^{(J)}(\Delta, x_0, x; \theta) b^{(J)}(y_0; \theta)\right\|_1 \\
&\leq \max_{0 \leq j \leq m-1} \left(y^{(j+1)} - y^{(j)} \right) \sum_{j=0}^{m-1} \sum_{k=0}^{m-1} \sup_{\varsigma \in (y^{(j)}, y^{(j+1)}]} \left| \frac{\partial \hat{B}_{(y^{(k)}, y^{(k+1)}),0}}{\partial y_0}(\Delta, x|x_0, \varsigma; \theta) \right| 1_{\{y^{(j)} < y_0 \leq y^{(j+1)}\}}.
\end{aligned}$$

By the definition of $\hat{B}_{(y^{(k)}, y^{(k+1)}),0}$, we have

$$\begin{aligned}
\sup_{\varsigma \in (y^{(j)}, y^{(j+1)}]} \left| \frac{\partial \hat{B}_{(y^{(k)}, y^{(k+1)}),0}}{\partial y_0}(\Delta, x|x_0, \varsigma; \theta) \right| &= \sup_{\varsigma \in (y^{(j)}, y^{(j+1)}]} \left| \frac{\partial}{\partial y_0} \int_{y^{(k)}}^{y^{(k+1)}} \hat{p}_{(X,Y)}(\Delta, x, y|x_0, \varsigma; \theta) dy \right| \\
&\leq \int_{y^{(k)}}^{y^{(k+1)}} \sup_{\varsigma \in (y^{(j)}, y^{(j+1)}]} \left| \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, \varsigma; \theta) \right| dy.
\end{aligned}$$

Since the first order derivative $\partial \hat{p}_{(X,Y)} / \partial y_0$ is uniformly continuous with respect to y_0 on the compact $\mathcal{Y}$, there exists $\delta > 0$, such that for any $\eta > 0$, as long as $y^{(j+1)} - y^{(j)} < \delta$, we have

$$\sup_{\varsigma \in (y^{(j)}, y^{(j+1)}]} \left| \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, \varsigma; \theta) \right| \leq \left| \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)}; \theta) \right| + \frac{\eta}{|\mathcal{Y}|}.$$

Here, $|\mathcal{Y}| < \infty$ represents the Lebesgue measure of the compact set $\mathcal{Y}$. Define the set $\mathcal{T}_{x,x_0,\theta}^{(k,j)}$ by

$$\mathcal{T}_{x,x_0,\theta}^{(k,j)} = \left\{ y \in (y^{(k)}, y^{(k+1)}] : \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)}; \theta) \geq 0 \right\}.$$

Then, we have

$$\sup_{\varsigma \in (y^{(j)}, y^{(j+1)}]} \left| \frac{\partial \hat{B}_{(y^{(k)}, y^{(k+1)}),0}}{\partial y_0}(\Delta, x|x_0, \varsigma; \theta) \right|$$

$$
\leq \int_{\mathcal{T}^{(k,j)}_{x,x_0,\theta}} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)};\theta)dy - \int_{(\mathcal{T}^{(k,j)}_{x,x_0,\theta})^c} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)};\theta)dy
$$
$$
+ \frac{\eta}{|\mathcal{Y}|}\int_{y^{(k)}}^{y^{(k+1)}} dy.
$$

By summing over k and j, we have

$$
\begin{aligned}
&\sum_{j=0}^{m-1}\sum_{k=0}^{m-1} \sup_{\varsigma\in(y^{(j)},y^{(j+1)}]} \left|\frac{\partial \hat{B}_{(y^{(k)},y^{(k+1)}),0}}{\partial y_0}(\Delta, x|x_0,\varsigma;\theta)\right| 1_{\{y^{(j)}<y_0\leq y^{(j+1)}\}} \\
&\leq \sum_{j=0}^{m-1} 1_{\{y^{(j)}<y_0\leq y^{(j+1)}\}} \left[\int_{\bigcup_{k=0}^{m-1}\mathcal{T}^{(k,j)}_{x,x_0,\theta}} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)};\theta)dy\right. \\
&\left. - \int_{\bigcup_{k=0}^{m-1}(\mathcal{T}^{(k,j)}_{x,x_0,\theta})^c} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)};\theta)dy + \eta\right] \\
&\leq \max_{0\leq j\leq m-1} \left[\int_{\bigcup_{k=0}^{m-1}\mathcal{T}^{(k,j)}_{x,x_0,\theta}} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)};\theta)dy\right. \\
&\left. - \int_{\bigcup_{k=0}^{m-1}(\mathcal{T}^{(k,j)}_{x,x_0,\theta})^c} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y^{(j)};\theta)dy + \eta\right].
\end{aligned}
$$

Similar to the discussions after Assumption 7 for the marginal transition density, we can show there exists $C_B > 0$, such that

$$
\sup_{0\leq k,j\leq m-1} \sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left|\int_{\bigcup_{k=0}^{m-1}\mathcal{T}^{(k,j)}_{x,x_0,\theta}} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y_0;\theta)dy\right| \leq C_B\Delta^{-\frac{1}{2}},
$$
$$
-\sup_{0\leq k,j\leq m-1} \sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left|\int_{\bigcup_{k=0}^{m-1}(\mathcal{T}^{(k,j)}_{x,x_0,\theta})^c} \frac{\partial \hat{p}_{(X,Y)}}{\partial y_0}(\Delta, x, y|x_0, y_0;\theta)dy\right| \leq C_B\Delta^{-\frac{1}{2}}.
$$

It follows that

$$
\left\|\hat{B}(\Delta, x|x_0, y_0;\theta) - \beta^{(J)}(\Delta, x_0, x;\theta)b^{(J)}(y_0;\theta)\right\|_1 \leq \max_{0\leq j\leq m-1}\left(y^{(j+1)} - y^{(j)}\right)\left(2C_B\Delta^{-\frac{1}{2}} + \eta\right).
$$

Taking supremum on both sides yields

$$
\begin{aligned}
&\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left\|\hat{B}(\Delta, x|x_0, y_0;\theta) - \beta^{(J)}(\Delta, x_0, x;\theta)b^{(J)}(y_0;\theta)\right\|_1 \\
&\leq \max_{0\leq j\leq m-1}\left(y^{(j+1)} - y^{(j)}\right)\left(2C_B\Delta^{-\frac{1}{2}} + \eta\right).
\end{aligned}
$$

For any $\epsilon > 0$, there exists a sufficiently large integer $\bar{m}^{\epsilon}_{\Delta,2} > 0$, such that for any $m > \bar{m}^{\epsilon}_{\Delta,2}$, we have

$$
\max_{0\leq k\leq m-1}\left(y^{(j+1)} - y^{(j)}\right) < \min\left\{\frac{\epsilon + \eta\Delta^{1/2}}{2C_B}, \delta\right\},
$$

and thus

$$
\sup_{(x_0,x,y_0,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta} \left\|\hat{B}(\Delta, x|x_0, y_0;\theta) - \beta^{(J)}(\Delta, x_0, x;\theta)b^{(J)}(y_0;\theta)\right\|_1 < \epsilon\Delta^{-\frac{1}{2}}.
$$

Assumption 9. *There exist positive constants K_1 and $\Delta_\alpha > 0$, such that for any $\Delta < \Delta_\alpha$ and $J > J_{\Delta,\alpha}$, the coefficient functions α_j satisfy*

$$\sup_{(x_0,x,\theta)\in\mathcal{X}^2\times\Theta}\left\|\alpha^{(J)}(\Delta,x_0,x;\theta)\right\|_1 \leq K_1\Delta^{-\frac{1}{2}}.$$

To justify Assumption 9 under the piecewise constant basis functions, recall the closed-form formula of $\alpha_{(y^{(k)},y^{(k+1)}),0}$ in (B.16), i.e.,

$$\alpha_{(y^{(k)},y^{(k+1)}),0}(\Delta,x_0,x;\theta) = \hat{p}_X(\Delta,x|x_0,y^{(k)};\theta),$$

where the closed-formula of $\hat{p}_X$ is given in (B.9). It follows from Assumptions 1 and 2 that $\hat{p}_X$ is bounded, i.e., there exists $K_1 > 0$, such that

$$\sup_{(x_0,x,\varsigma,\theta)\in\mathcal{X}^2\times\mathcal{Y}\times\Theta}|\hat{p}_X(\Delta,x|x_0,\varsigma;\theta)| \leq K_1\Delta^{-\frac{1}{2}}.$$

It follows from the definition of L_1-norm that

$$\sup_{(x_0,x,\theta)\in\mathcal{X}^2\times\Theta}\left\|\alpha^{(J)}(\Delta,x_0,x;\theta)\right\|_1 = \sup_{(x_0,x,\theta)\in\mathcal{X}^2\times\Theta}\max_{0\leq k\leq m-1}\left|\hat{p}_X(\Delta,x|x_0,y^{(k)};\theta)\right| \leq K_1\Delta^{-\frac{1}{2}}.$$

Assumption 10. *There exist positive constants K_2 and $\Delta_\beta > 0$, such that for any $\Delta < \Delta_\beta$ and $J > J_{\Delta,\beta}$, the coefficient functions $\beta_{i,j}$ satisfy*

$$\sup_{(x_0,x,\theta)\in\mathcal{X}^2\times\Theta}\left\|\beta^{(J)}(\Delta,x_0,x;\theta)\right\|_1 \leq K_2\Delta^{-\frac{1}{2}}. \tag{B.18}$$

To justify Assumption 10 under the piecewise constant basis functions, recall the closed-form formula of $\beta_{(y^{(k)},y^{(k+1)}),0}^{(y^{(j)},y^{(j+1)}),0}$ in (B.17), i.e.,

$$\beta_{(y^{(k)},y^{(k+1)}),0}^{(y^{(j)},y^{(j+1)}),0}(\Delta,x_0,x;\theta) = \hat{B}_{(y^{(k)},y^{(k+1)}),0}(\Delta,x|x_0,y^{(j)};\theta).$$

It follows from the definition of L_1-norm that

$$\begin{aligned}\left\|\beta^{(J)}(\Delta,x_0,x;\theta)\right\|_1 &= \max_{0\leq j\leq m-1}\sum_{k=0}^{m-1}\left|\hat{B}_{(y^{(k)},y^{(k+1)}),0}(\Delta,x|x_0,y^{(j)};\theta)\right| \\ &= \max_{0\leq j\leq m-1}\sum_{k=0}^{m-1}\left|\int_{y^{(k)}}^{y^{(k+1)}}\hat{p}_{(X,Y)}(\Delta,x,y|x_0,y^{(j)};\theta)dy\right|.\end{aligned}$$

Without loss of generality, we assume $y^{(0)} \leq 0$ and $y^{(m)} \geq 0$. Then, we obtain

$$\begin{aligned}\left\|\beta^{(J)}(\Delta,x_0,x;\theta)\right\|_1 \leq &-\max_{0\leq j\leq m-1}\int_{y^{(0)}}^{0}\hat{p}_{(X,Y)}(\Delta,x,y|x_0,y^{(j)};\theta)dy \\ &+\max_{0\leq j\leq m-1}\int_{0}^{y^{(m)}}\hat{p}_{(X,Y)}(\Delta,x,y|x_0,y^{(j)};\theta)dy\end{aligned}$$

Similar to the discussions after Assumption 7 for the marginal transition density, we can show there exists $\bar{C}_p > 0$, such that

$$-\sup_{(x_0,x,\theta)\in\mathcal{X}^2\times\Theta}\max_{0\leq j\leq m-1}\int_{y^{(0)}}^{0}\hat{p}_{(X,Y)}(\Delta,x,y|x_0,y^{(j)};\theta)dy \leq \bar{C}_p\Delta^{-\frac{1}{2}},$$

$$\sup_{(x_0,x,\theta)\in\mathcal{X}^2\times\Theta}\max_{0\leq j\leq m-1}\int_{0}^{y^{(m)}}\hat{p}_{(X,Y)}(\Delta,x,y|x_0,y^{(j)};\theta)dy \leq \bar{C}_p\Delta^{-\frac{1}{2}}.$$

Finally, (B.18) follows from denoting $K_2 = 2\bar{C}_p$.

Appendix C Proofs

Appendix C.1 Proof of Corollary 1

Proof. Recall definition (18) of the conditional generalized moment $\mathcal{G}_i(\theta)$, i.e.,

$$\mathcal{G}_i(\theta) = \mathbb{E}[g(X_{i\Delta};\theta)|\mathbf{X}_{(i-1)\Delta};\theta] = \int_{\mathcal{X}} g(x_{i\Delta};\theta)p(x_{i\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dx_{i\Delta}.$$

By the conditional probability formula, we have

$$\begin{aligned}\mathcal{G}_i(\theta) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} g(x_{i\Delta};\theta)p(x_{i\Delta},y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta})dy_{(i-1)\Delta}dx_{i\Delta}\\ &= \int_{\mathcal{Y}}\int_{\mathcal{X}} g(x_{i\Delta};\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)p(x_{i\Delta}|\mathbf{X}_{(i-1)\Delta},y_{(i-1)\Delta};\theta)dx_{i\Delta}dy_{(i-1)\Delta}.\end{aligned}$$

Thanks to the Markov property of the model (1a)–(1b), the conditional density $p(x_{i\Delta}|\mathbf{X}_{(i-1)\Delta},y_{(i-1)\Delta};\theta)$ coincides with the marginal transition density $p_X(\Delta,x_{i\Delta}|X_{(i-1)\Delta},y_{(i-1)\Delta};\theta)$. As a result, we have

$$\mathcal{G}_i(\theta) = \int_{\mathcal{Y}} G(\Delta,X_{(i-1)\Delta},y_{(i-1)\Delta};\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}.$$

Plugging the assumed approximation (19) into the right hand side, we obtain the following approximation

$$\tilde{\mathcal{G}}_i^{(J)}(\theta) = \sum_{k=0}^{J}\gamma_k\left(X_{(i-1)\Delta};\theta\right)\mathcal{M}_{k,(i-1)\Delta}(\theta).$$

Then, (20) follows from further replacing the generalized filtered moments $\mathcal{M}_{k,(i-1)\Delta}$ by their approximations $\hat{\mathcal{M}}_{k,(i-1)\Delta}^{(J)}$ on the right hand side. □

Appendix C.2 Proof of Theorem 2

As a preparation, we propose the following proposition before proving Theorem 2.

Proposition 2. *Under Assumptions 1–10 provided in Appendix B, for any number of observations n, the approximations of the generalized filtered moments $\hat{\mathcal{M}}_{k,i\Delta}^{(J)}(\Delta,\theta)$, likelihood update $\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)$,* *log-likelihood update $\hat{\ell}_i^{(J)}(\Delta,\theta)$, and log-likelihood $\hat{\ell}(\Delta,\theta)$ converge to the corresponding true versions in the following sense: for any $\epsilon > 0$, there exists $\Delta^\epsilon > 0$, such that for any $\Delta < \Delta^\epsilon$ and $J \geq J_\Delta^\epsilon$, we have*

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{(i-1)\Delta}(\Delta,\theta)\right\|_1 \leq \rho_M^{(i-1)}(\Delta,\epsilon)\right) \geq 1-\epsilon, \tag{C.1}$$

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\frac{\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)-\mathcal{L}_i(\Delta,\theta)}{\mathcal{L}_i(\Delta,\theta)}\right| \leq \rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)\right) \geq 1-\epsilon, \tag{C.2}$$

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right| \leq \rho_\ell^{(i)}(\Delta,\epsilon)\right) \geq 1-\epsilon, \tag{C.3}$$

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\hat{\ell}^{(J)}(\Delta,\theta)-\ell(\Delta,\theta)\right| \leq \rho^{(n)}(\Delta,\epsilon)\right) \geq 1-\epsilon, \tag{C.4}$$

for $i = 1,2,\ldots,n$. Here, J_Δ^ϵ is a function of Δ and ϵ; the upper bounds of errors $\rho_M^{(i-1)}(\Delta,\epsilon)$, $\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)$, $\rho_\ell^{(i)}(\Delta,\epsilon)$, and $\rho^{(n)}(\Delta,\epsilon)$ converge to zero as $\Delta \to 0$. Moreover, for any $\epsilon,\eta > 0$, there exist sequences $\Delta_n^{\epsilon,\eta}$ and $J_n^{\epsilon,\eta}$, such that as $n \to \infty$, we have $\Delta_n^{\epsilon,\eta} \to 0$, $J_n^{\epsilon,\eta} \to \infty$, and

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\hat{\ell}^{(J_n^{\epsilon,\eta})}(\Delta_n^{\epsilon,\eta},\theta)-\ell(\Delta_n^{\epsilon,\eta},\theta)\right| \leq \eta\right) \geq 1-\epsilon. \tag{C.5}$$

Proof. See Appendix C.2.1. □

Now, we prove Theorem 2.

Proof. To prove the four convergence formulae in (35) and (36), without loss of generality, we take the second one in (36) as an example. For any $\epsilon > 0$, since the function $\rho^{(n)}(\Delta,\epsilon) \to 0$ as $\Delta \to 0$, we have for any $\eta > 0$, there exists $\bar{\Delta}^{\epsilon,\eta}$, such that for any $\Delta < \bar{\Delta}^{\epsilon,\eta}$,

$$\rho^{(n)}(\Delta,\epsilon) < \eta.$$

Combining the above inequality with (C.4), we obtain for any $\epsilon,\eta > 0$, $\Delta < \bar{\Delta}^{\epsilon,\eta}$, and $J > \bar{J}_\Delta^\epsilon$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\hat{\ell}^{(J)}(\Delta,\theta)-\ell(\Delta,\theta)\right| > \eta\right) < \epsilon.$$

Here, $\bar{J}_\Delta^\epsilon$ is defined by $\bar{J}_\Delta^\epsilon = \max\{J_\Delta^\epsilon, 1/\Delta\}$, where J_Δ^ϵ is employed in Proposition 2. By definition, we note that $\bar{J}_\Delta^\epsilon \to \infty$ as $\Delta \to 0$ for any $\epsilon > 0$. This immediately implies the desired result.

For the convergence formula of AMMLE (37) to MMLE follows from standard arguments based on the convergence of the log-likelihood function stated in (36); see, e.g., the proofs for Theorem 3 of Aït-Sahalia (2008) and Proposition 1 of Li (2013) among others. From the assumed existence of the maximizer $\hat{\theta}_{\mathrm{MMLE}}^{(n,\Delta)}$ of $\ell(\Delta,\theta)$ in Θ and the above justified proximity of the two objective functions $\ell(\Delta,\theta)$ and $\hat{\ell}^{(J)}(\Delta,\theta)$, the maximizer of $\hat{\ell}^{(J)}(\Delta,\theta)$, that is $\hat{\theta}_{\mathrm{AMMLE}}^{(n,\Delta,J)}$, exists in Θ with probability approaching to one as $\Delta \to 0$ and $J \to \infty$. Moreover, these two maximizers are close to each other in the sense of (37).

Finally, we prove the convergence formulae in (38) as $n \to \infty$. The first statement in (38) is mathematically equivalent to (C.5) in Proposition 2. It simply rewrites (C.5) as a limit instead of using $\epsilon - \delta$ language. Similar to the arguments linking (36) to (37), the second statement in (38) is a natural implication of the first one. □

Appendix C.2.1 Proof of Proposition 2

Proof. The proof includes three steps. In the first step, we prove (C.1), which also serves as a foundation of the rest of the statements in Proposition 2. In the second step, we prove (C.2)–(C.4). Finally, in the third step, we prove (C.5).

Step 1 – The convergence of the generalized filtered moment approximations. We will prove (C.1) by mathematical induction. At the initial stage, we have by construction $\mathcal{M}_{k,0}(\Delta,\theta) = \hat{\mathcal{M}}_{k,0}^{(J)}(\Delta,\theta)$. This immediately implies

$$\sup_{\theta\in\Theta}\left\|\mathcal{M}_0(\Delta,\theta) - \hat{\mathcal{M}}_0^{(J)}(\Delta,\theta)\right\|_1 = \sup_{\theta\in\Theta}\sum_{k=0}^{J}\left|\mathcal{M}_{k,0}(\Delta,\theta) - \hat{\mathcal{M}}_{k,0}^{(J)}(\Delta,\theta)\right| = 0,$$

which corresponds to (C.1) with $\rho_M^{(0)}(\Delta,\epsilon) \equiv 0$ and $J_\Delta^{\epsilon,0} = 0$ for any $\Delta > 0$. Now, suppose (C.1) holds at the $(i-1)$th stage, i.e., there exists $\Delta^{\epsilon,(i-1)} > 0$, functions $J_\Delta^{\epsilon,(i-1)}$ and $\rho_M^{(i-1)}(\Delta,\epsilon)$ such that for any $\Delta < \Delta^{\epsilon,(i-1)}$ and $J > J_\Delta^{\epsilon,(i-1)}$, we have

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta) - \mathcal{M}_{(i-1)\Delta}(\Delta,\theta)\right\|_1 \le \rho_M^{(i-1)}(\Delta,\epsilon)\right) \ge 1-\epsilon, \tag{C.6}$$

and $\rho_M^{(i-1)}(\Delta,\epsilon) \to 0$ as $\Delta \to 0$. We will prove (C.1) is true at the ith stage and $\rho_M^{(i)}(\Delta,\epsilon) \to 0$, as $\Delta \to 0$.

By definition, we have

$$\begin{aligned}\left\|\mathcal{M}_{i\Delta}(\Delta,\theta) - \hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)\right\|_1 = \Bigg\| &\frac{\int_{\mathcal{Y}} B(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}}{\mathcal{L}_i(\Delta,\theta)} \\ &- \frac{\beta^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta})\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)}{\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)}\Bigg\|_1,\end{aligned}$$

It follows from the triangle inequality that

$$\left\|\mathcal{M}_{i\Delta}(\Delta,\theta) - \hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)\right\|_1 \le R_1(\Delta,\theta) + R_2(\Delta,\theta),$$

where

$$\begin{aligned}R_1(\Delta,\theta) = \frac{1}{\mathcal{L}_i(\Delta,\theta)}\Bigg\|&\int_{\mathcal{Y}} B(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta} \\ &-\beta^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta};\theta)\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\Bigg\|_1,\end{aligned} \tag{C.7a}$$

and

$$R_2(\Delta,\theta) = \frac{1}{\mathcal{L}_i(\Delta,\theta)\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)} \left\|\beta^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta};\theta)\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1$$
$$\times \left|\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta) - \mathcal{L}_i(\Delta,\theta)\right|. \tag{C.7b}$$

To deal with the first term $R_1(\Delta,\theta)$ in (C.7a), we note that the closed-form formulae of the marginal transition moments B are generally implicit. Thus, according to our algorithm, we need to introduce an approximation $\hat{B}$, e.g., the Euler approximation of B, to calculate the coefficient matrix β. By employing $\hat{B} = (\hat{B}_0, \hat{B}_1, \cdots, \hat{B}_J)^\intercal$, the triangle inequality implies

$$R_1(\Delta,\theta) \leq R_1^{(1)}(\Delta,\theta) + R_1^{(2)}(\Delta,\theta) + R_1^{(3)}(\Delta,\theta),$$

where

$$R_1^{(1)}(\Delta,\theta) = \frac{1}{\mathcal{L}_i(\Delta,\theta)} \left\| \int_{\mathcal{Y}} \left[B(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta) - \hat{B}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)\right]\right.$$
$$\left. \times p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}\right\|_1,$$
$$R_1^{(2)}(\Delta,\theta) = \frac{1}{\mathcal{L}_i(\Delta,\theta)} \left\| \int_{\mathcal{Y}} \hat{B}(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta)p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}\right.$$
$$\left. -\beta^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta};\theta)\mathcal{M}_{(i-1)\Delta}(\Delta,\theta)\right\|_1,$$
$$R_1^{(3)}(\Delta,\theta) = \frac{1}{\mathcal{L}_i(\Delta,\theta)} \left\|\beta^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta};\theta)\left[\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right]\right\|_1.$$

To prove (C.1), note that for any $\Delta < \Delta_{\mathcal{L}}^{\epsilon/3}$, we have

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta) - \mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq \rho_M^{(i)}(\Delta,\epsilon)\right)$$
$$\geq \mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta) - \mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq \rho_M^{(i)}(\Delta,\epsilon), \inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta) \geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}\right)$$
$$= \mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta) - \mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq \rho_M^{(i)}(\Delta,\epsilon)\middle| \inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta) \geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}\right)$$
$$\times \mathbb{P}\left(\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta) \geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}\right)$$
$$\geq \left(1-\frac{\epsilon}{3}\right)\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta) - \mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq \rho_M^{(i)}(\Delta,\epsilon)\middle| \inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta) \geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}\right), \tag{C.8}$$

where the last inequality follows from Lemma 2. Thus, it suffices to find the upper bound $\rho_M^{(i)}(\Delta,\epsilon)$ such that the last conditional probability can be sufficiently close to 1. We develop all the rest of arguments given that $\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta) \geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-1/2}$.

For $R_1^{(1)}(\Delta,\theta)$, it follows from Assumption 6 that

$$R_1^{(1)}(\Delta,\theta) \leq \frac{1}{\mathcal{L}_i(\Delta,\theta)} \int_{\mathcal{Y}} \left\|\left[B(\Delta, X_{(i-1)\Delta}|X_{i\Delta}, y_{(i-1)\Delta};\theta) - \hat{B}(\Delta, X_{(i-1)\Delta}|X_{i\Delta}, y_{(i-1)\Delta};\theta)\right]\right\|_1$$

$$\begin{aligned}&\times p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}\\ &\leq \frac{1}{\mathcal{L}_i(\Delta,\theta)}\rho_2(\Delta)\Delta^{-\frac{1}{2}}.\end{aligned}$$

Combining the above inequality and (B.6), we have

$$R_1^{(1)}(\Delta,\theta) \leq \frac{1}{C_{\mathcal{L}}^{\epsilon/3}}\rho_2(\Delta). \tag{C.9}$$

For $R_1^{(2)}(\Delta,\theta)$, by plugging in the definition of the generalized filtered moments $\mathcal{M}_{(i-1)\Delta}$, we obtain

$$\begin{aligned}R_1^{(2)}(\Delta,\theta) = \frac{1}{\mathcal{L}_i(\Delta,\theta)}&\left\|\int_{\mathcal{Y}}\left[\hat{B}(\Delta,X_{(i-1)\Delta}|X_{i\Delta},y_{(i-1)\Delta};\theta) - \beta^{(J)}(\Delta,X_{(i-1)\Delta},X_{i\Delta};\theta)b^{(J)}(y_{(i-1)\Delta};\theta)\right]\right.\\ &\left.\times p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}\right\|_1.\end{aligned}$$

According to Assumption 8, for any $\Delta > 0$, there exists $J_{\Delta,2} > 0$, such that for any $J \geq J_{\Delta,2}$, we have

$$\begin{aligned}R_1^{(2)}(\Delta,\theta) \leq \frac{1}{\mathcal{L}_i(\Delta,\theta)}&\int_{\mathcal{Y}}\left\|\left[\hat{B}(\Delta,X_{(i-1)\Delta}|X_{i\Delta},y_{(i-1)\Delta};\theta) - \beta^{(J)}(\Delta,X_{(i-1)\Delta},X_{i\Delta};\theta)b^{(J)}(y_{(i-1)\Delta};\theta)\right]\right\|_1\\ &\times p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta)dy_{(i-1)\Delta}\\ &\leq \frac{1}{\mathcal{L}_i(\Delta,\theta)}\eta_2(\Delta,J_{\Delta,2})\Delta^{-\frac{1}{2}}.\end{aligned}$$

Combining the above inequality and (B.6), we have

$$R_1^{(2)}(\Delta,\theta) \leq \frac{1}{C_{\mathcal{L}}^{\epsilon/3}}\eta_2(\Delta,J_{\Delta,2}). \tag{C.10}$$

For $R_1^{(3)}(\Delta,\theta)$, note that

$$R_1^{(3)}(\Delta,\theta) \leq \frac{1}{C_{\mathcal{L}}^{\epsilon/3}}\Delta^{\frac{1}{2}}\left\|\beta^{(J)}(\Delta,X_{(i-1)\Delta},X_{i\Delta};\theta)\right\|_1\left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1.$$

According to Assumption 10, there exists a constant $\Delta_\beta > 0$, such that for any $\Delta < \Delta_\beta$ and $J > J_{\Delta,\beta}$, we have

$$R_1^{(3)}(\Delta,\theta) \leq \frac{K_2}{C_{\mathcal{L}}^{\epsilon/3}}\left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1. \tag{C.11}$$

Combining (C.9), (C.10), and (C.11), we obtain

$$R_1(\Delta,\theta) \leq \frac{1}{C_{\mathcal{L}}^{\epsilon/3}}\left(\rho_2(\Delta) + \eta_2(\Delta,J_{\Delta,2})\right) + \frac{K_2}{C_{\mathcal{L}}^{\epsilon/3}}\left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1. \tag{C.12}$$

On the other hand, to deal with the second term $R_2(\Delta,\theta)$ in (C.7b), we first prove the boundedness of the second multiplier $\left\|\beta(\Delta,X_{(i-1)\Delta},X_{i\Delta};\theta)\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1$. Indeed, we note that

$$\left\|\beta^{(J)}(\Delta,X_{(i-1)\Delta},X_{i\Delta};\theta)\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1 \leq \left\|\beta^{(J)}(\Delta,X_{(i-1)\Delta},X_{i\Delta};\theta)\right\|_1\left\|\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1$$

$$\leq K_2 \Delta^{-\frac{1}{2}} \left\| \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \right\|_1,$$

where the second inequality follows from Assumption 10. To bound the L_1-norm of approximate generalized filtered moments $\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}$, the triangle inequality suggests

$$\begin{aligned}\left\| \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \right\|_1 &\leq \left\| \mathcal{M}_{(i-1)\Delta}(\Delta, \theta) \right\|_1 + \left\| \mathcal{M}_{(i-1)\Delta}(\Delta, \theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \right\|_1. \\ &\leq C_0 + \left\| \mathcal{M}_{(i-1)\Delta}(\Delta, \theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \right\|_1,\end{aligned}$$

where the second inequality is implied by Assumption 4. As a consequence, we have

$$\begin{aligned}&\left\| \beta^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta) \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \right\|_1 \\ &\leq K_2 \Delta^{-\frac{1}{2}} \left(C_0 + \left\| \mathcal{M}_{(i-1)\Delta}(\Delta, \theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \right\|_1 \right).\end{aligned} \tag{C.13}$$

Next, we inspect the error of likelihood update approximation $\left| \hat{\mathcal{L}}_i^{(J)}(\Delta, \theta) - \mathcal{L}_i(\Delta, \theta) \right|$ in $R_2(\Delta, \theta)$. By definition, we have

$$\begin{aligned}\left| \hat{\mathcal{L}}_i^{(J)}(\Delta, \theta) - \mathcal{L}_i(\Delta, \theta) \right| = \bigg| &\int_{\mathcal{Y}} p_X(\Delta, X_{i\Delta} | X_{(i-1)\Delta}, y_{(i-1)\Delta}; \theta) p(y_{(i-1)\Delta} | \mathbf{X}_{(i-1)\Delta}; \theta) dy_{(i-1)\Delta} \\ &- \alpha^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta) \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \bigg|.\end{aligned}$$

It follows from the triangle inequality that

$$\left| \hat{\mathcal{L}}_i^{(J)}(\Delta, \theta) - \mathcal{L}_i(\Delta, \theta) \right| \leq S_1(\Delta, \theta) + S_2(\Delta, \theta) + S_3(\Delta, \theta),$$

where

$$\begin{aligned}S_1(\Delta, \theta) &= \bigg| \int_{\mathcal{Y}} \left[p_X(\Delta, X_{i\Delta} | X_{(i-1)\Delta}, y_{(i-1)\Delta}; \theta) - \hat{p}_X(\Delta, X_{i\Delta} | X_{(i-1)\Delta}, y_{(i-1)\Delta}; \theta) \right] \\ &\quad \times p(y_{(i-1)\Delta} | \mathbf{X}_{(i-1)\Delta}; \theta) dy_{(i-1)\Delta} \bigg|, \\ S_2(\Delta, \theta) &= \bigg| \int_{\mathcal{Y}} \hat{p}_X(\Delta, X_{i\Delta} | X_{(i-1)\Delta}, y_{(i-1)\Delta}; \theta) p(y_{(i-1)\Delta} | \mathbf{X}_{(i-1)\Delta}; \theta) dy_{(i-1)\Delta} \\ &\quad - \alpha^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta) \mathcal{M}_{(i-1)\Delta}(\Delta, \theta) \bigg|, \\ S_3(\Delta, \theta) &= \left| \alpha^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta) \left[\mathcal{M}_{(i-1)\Delta}(\Delta, \theta) - \hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta, \theta) \right] \right|.\end{aligned}$$

For $S_1(\Delta, \theta)$, Assumption 5 implies

$$\begin{aligned}S_1(\Delta, \theta) &\leq \int_{\mathcal{Y}} \left| p_X(\Delta, X_{i\Delta} | X_{(i-1)\Delta}, y_{(i-1)\Delta}; \theta) - \hat{p}_X(\Delta, X_{i\Delta} | X_{(i-1)\Delta}, y_{(i-1)\Delta}; \theta) \right| \\ &\quad \times p(y_{(i-1)\Delta} | \mathbf{X}_{(i-1)\Delta}; \theta) dy_{(i-1)\Delta} \\ &\leq \rho_1(\Delta) \Delta^{-\frac{1}{2}}.\end{aligned} \tag{C.14}$$

For $S_2(\Delta,\theta)$, plugging in the definition of the generalized filtered moments $\mathcal{M}_{(i-1)\Delta}$, we obtain

$$S_2(\Delta,\theta) = \bigg|\int_{\mathcal{Y}} \Big[\hat{p}_X(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta) - \alpha^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta};\theta) b^{(J)}(y_{(i-1)\Delta};\theta)\Big] \\ \times p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta) dy_{(i-1)\Delta}\bigg|.$$

According to Assumption 7, there exists a $J_{\Delta,1}$, such that for any $J > J_{\Delta,1}$, we have

$$\begin{aligned} S_2(\Delta,\theta) \leq & \int_{\mathcal{Y}} \left|\hat{p}_X(\Delta, X_{i\Delta}|X_{(i-1)\Delta}, y_{(i-1)\Delta};\theta) - \alpha^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta};\theta) b^{(J)}(y_{(i-1)\Delta};\theta)\right| \\ & \times p(y_{(i-1)\Delta}|\mathbf{X}_{(i-1)\Delta};\theta) dy_{(i-1)\Delta} \\ \leq & \, \eta_1(\Delta, J_{\Delta,1})\Delta^{-\frac{1}{2}}. \end{aligned} \tag{C.15}$$

For $S_3(\Delta,\theta)$, we note that

$$\begin{aligned} S_3(\Delta,\theta) &\leq \left\|\alpha^{(J)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta};\theta)\right\|_1 \left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}^{(J)}_{(i-1)\Delta}(\Delta,\theta)\right\|_1 \\ &\leq K_1\Delta^{-\frac{1}{2}} \left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}^{(J)}_{(i-1)\Delta}(\Delta,\theta)\right\|_1, \end{aligned} \tag{C.16}$$

where the second inequality is implied by Assumption 9. Combining (C.14), (C.15), and (C.16), we arrive at

$$\begin{aligned} &\left|\hat{\mathcal{L}}^{(J)}_i(\Delta,\theta) - \mathcal{L}_i(\Delta,\theta)\right| \\ &\leq \Delta^{-\frac{1}{2}} \left[\rho_1(\Delta) + \eta_1(\Delta, J_{\Delta,1}) + K_1 \left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}^{(J)}_{(i-1)\Delta}(\Delta,\theta)\right\|_1\right]. \end{aligned} \tag{C.17}$$

Finally, we investigate the first multiplier $1/(\mathcal{L}_i(\Delta,\theta)\hat{\mathcal{L}}^{(J)}_i(\Delta,\theta))$ in $R_2(\Delta,\theta)$. According to (C.17), we have

$$\begin{aligned} \left|\hat{\mathcal{L}}^{(J)}_i(\Delta,\theta)\right| &\geq \mathcal{L}_i(\Delta,\theta) - \Delta^{-\frac{1}{2}} \left[\rho_1(\Delta) + \eta_1(\Delta, J_{\Delta,1}) + K_1 \left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}^{(J)}_{(i-1)\Delta}(\Delta,\theta)\right\|_1\right] \\ &\geq \Delta^{-\frac{1}{2}} \left(C^{\epsilon/3}_{\mathcal{L}} - K_1 \left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}^{(J)}_{(i-1)\Delta}(\Delta,\theta)\right\|_1 - \rho_1(\Delta) - \eta_1(\Delta, J_{\Delta,1})\right), \end{aligned}$$

where the second inequality follows by (B.6). By the assumptions of our mathematical induction (C.6), for any $\epsilon/3 > 0$, there exists $\Delta_1^{\epsilon/3} > 0$, such that for any $\Delta < \Delta_1^{\epsilon/3}$ and $J \geq \bar{J}^{\epsilon/3}_{\Delta}$, we have $\mathbb{P}(A_i) > 1 - \epsilon/3$, where A_i is the event given by

$$A_i = \left\{K_1 \sup_{\theta\in\Theta} \left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta) - \hat{\mathcal{M}}^{(J)}_{(i-1)\Delta}(\Delta,\theta)\right\|_1 + \rho_1(\Delta) + \eta_1(\Delta, J_{\Delta,1}) \leq \frac{C^{\epsilon/3}_{\mathcal{L}}}{2}\right\}.$$

Then, the probability inequality (C.8) can be further revised as

$$\begin{aligned} &\mathbb{P}\left(\sup_{\theta\in\Theta} \left\|\hat{\mathcal{M}}^{(J)}_{i\Delta}(\Delta,\theta) - \mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq \rho^{(i)}_M(\Delta,\epsilon)\right) \\ &\geq \mathbb{P}\left(\sup_{\theta\in\Theta} \left\|\hat{\mathcal{M}}^{(J)}_{i\Delta}(\Delta,\theta) - \mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq \rho^{(i)}_M(\Delta,\epsilon), \inf_{\theta\in\Theta} \mathcal{L}_i(\Delta,\theta) \geq C^{\epsilon/3}_{\mathcal{L}}\Delta^{-\frac{1}{2}}, A_i\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \le \rho_M^{(i)}(\Delta,\epsilon)\middle| \inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\ge C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}, A_i\right)\\
&\times \mathbb{P}\left(\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\ge C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}, A_i\right),
\end{aligned}
$$

Thanks to the fact that

$$
\begin{aligned}
\mathbb{P}\left(\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\ge C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}, A_i\right) &= 1-\mathbb{P}\left(\left\{\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)< C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}\right\}\cup A_i^c\right)\\
&\ge 1-\mathbb{P}\left(\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)< C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}\right)-\mathbb{P}\left(A_i^c\right)\\
&\ge 1-\frac{2\epsilon}{3},
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \le \rho_M^{(i)}(\Delta,\epsilon)\right)\\
&\ge \mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \le \rho_M^{(i)}(\Delta,\epsilon)\middle| \inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\ge C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}, A_i\right)\\
&\times\left(1-\frac{2\epsilon}{3}\right). \qquad \text{(C.18)}
\end{aligned}
$$

Thus, similar to the discussion around (C.8), it suffices to find an appropriate upper bound $\rho_M^{(i)}(\Delta,\epsilon)$, such that the above conditional probability is close to 1. Starting from here, all the rest of arguments will be conducted given the occurrence of A_i and $\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\ge C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}$. Given these two events, it turns out that

$$
\left|\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)\right| \ge \frac{1}{2}C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}},
$$

and thus

$$
\frac{1}{\mathcal{L}_i(\Delta,\theta)\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)} \le \frac{2\Delta}{(C_{\mathcal{L}}^{\epsilon/3})^2}. \qquad \text{(C.19)}
$$

Combining (C.13), (C.17), and (C.19), we obtain

$$
\begin{aligned}
R_2(\Delta,\theta) \le \frac{2K_2}{(C_{\mathcal{L}}^{\epsilon/3})^2}&\left(C_0+\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1\right)\\
&\times\left[\rho_1(\Delta)+\eta_1(\Delta,J_{\Delta,1})+K_1\left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta)-\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1\right]. \qquad \text{(C.20)}
\end{aligned}
$$

By the assumption of our mathematical induction (C.6), for any $\epsilon/3>0$, there exists $\Delta^{\epsilon/3,(i-1)}>0$, such that for any $\Delta<\Delta^{\epsilon/3,(i-1)}$ and $J\ge J_\Delta^{\epsilon/3,(i-1)}$, we have $\mathbb{P}(A_i')>1-\epsilon/3$, where A_i' is the event given by

$$
A_i'=\left\{\sup_{\theta\in\Theta}\left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta)-\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1<\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\right\}.
$$

Similar to the probability inequalities (C.8) and (C.18), we derive

$$
\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \le \rho_M^{(i)}(\Delta,\epsilon)\right)
$$

$$\geq (1-\epsilon)\,\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)-\mathcal{M}_{i\Delta}(\Delta,\theta)\right\|_1 \leq \rho_M^{(i)}(\Delta,\epsilon)\middle|\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}},A_i,A_i'\right).$$

Now, it suffices to find an appropriate $\rho_M^{(i)}(\Delta,\epsilon)$ such that the above conditional probability is 1. Given the occurrence of A_i, A_i', and $\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}$, we further combine (C.12) and (C.20), take supremum with respect to θ, and finally obtain

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\mathcal{M}_{i\Delta}(\Delta,\theta)-\hat{\mathcal{M}}_{i\Delta}^{(J)}(\Delta,\theta)\right\|_1\\
&\leq \frac{1}{C_{\mathcal{L}}^{\epsilon/3}}\left(\rho_2(\Delta)+\eta_2(\Delta,J_{\Delta,2})\right)+\frac{K_2}{C_{\mathcal{L}}^{\epsilon/3}}\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\\
&+\frac{2K_2}{(C_{\mathcal{L}}^{\epsilon/3})^2}\left(C_0+\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\right)\left[\rho_1(\Delta)\Delta^{\frac{1}{2}}+\eta_1(\Delta,J_{\Delta,1})\Delta^{\frac{1}{2}}+K_1\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\right].
\end{aligned}$$

Hence, (C.1) is true at the ith stage after introducing

$$\begin{aligned}
\rho_M^{(i)}(\Delta,\epsilon)=&\frac{1}{C_{\mathcal{L}}^{\epsilon/3}}\Delta^{\frac{1}{2}}\left(\rho_2(\Delta)+\eta_2(\Delta,J_{\Delta,2})\right)+\frac{K_2}{C_{\mathcal{L}}^{\epsilon/3}}\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\\
&+\frac{2K_2}{(C_{\mathcal{L}}^{\epsilon/3})^2}\left(C_0+\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\right)\left[\rho_1(\Delta)+\eta_1(\Delta,J_{\Delta,1})+K_1\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\right].
\end{aligned}$$

Apparently, by the definition of $\rho_M^{(i)}(\Delta,\epsilon)$, we obtain $\rho_M^{(i)}(\Delta,\epsilon)\to 0$ as $\Delta\to 0$.

Tracing back all the above discussion, at the ith stage, we recursively define

$$\Delta^{\epsilon,(i)}=\min\{\Delta_{\mathcal{L}}^{\epsilon/3},\Delta_\alpha,\Delta_\beta,\Delta_1^{\epsilon/3},\Delta^{\epsilon/3,(i-1)}\}\text{ and }J_\Delta^{\epsilon,(i)}=\max\{J_{\Delta,1},J_{\Delta,2},J_\Delta^{\epsilon/3,(i-1)},\bar{J}_\Delta^{\epsilon/3},J_{\Delta,\alpha},J_{\Delta,\beta}\},$$

for $i=1,2,\ldots,n$. Eventually, we choose

$$\Delta^\epsilon=\min_{i=1,2,\ldots,n}\Delta^{\epsilon,(i)}\text{ and }J_\Delta^\epsilon=\max_{i=1,2,\ldots,n}J_\Delta^{\epsilon,(i)},$$

so that for any $\Delta<\Delta^\epsilon$, $J>J_\Delta^\epsilon$, and $i=1,2,\ldots,n$, (C.1) holds.

Step 2 – The convergence of the approximations for the likelihood update, log-likelihood update, and log-likelihood. First, as to the approximation error of the likelihood update, recall (C.17):

$$\left|\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)-\mathcal{L}_i(\Delta,\theta)\right|\leq\Delta^{-\frac{1}{2}}\left[\rho_1(\Delta)+\eta_1(\Delta,J_{\Delta,1})+K_1\left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta)-\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1\right],$$

which implies

$$\begin{aligned}
\left|\frac{\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)-\mathcal{L}_i(\Delta,\theta)}{\mathcal{L}_i(\Delta,\theta)}\right|\leq&\frac{1}{\mathcal{L}_i(\Delta,\theta)}\Delta^{-\frac{1}{2}}\left[K_1\left\|\mathcal{M}_{(i-1)\Delta}(\Delta,\theta)-\hat{\mathcal{M}}_{(i-1)\Delta}^{(J)}(\Delta,\theta)\right\|_1\right.\\
&\left.+\rho_1(\Delta)+\eta_1(\Delta,J_{\Delta,1})\right].
\end{aligned}$$

Given the occurrence of A_i, A_i', and $\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}}$, we take supremum on both sides and obtain

$$\sup_{\theta\in\Theta}\left|\frac{\mathcal{L}_i(\Delta,\theta)-\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)}{\mathcal{L}_i(\Delta,\theta)}\right|\leq\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon),$$

where

$$\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)=\frac{1}{C_{\mathcal{L}}^{\epsilon/3}}\left[\rho_1(\Delta)+\eta_1(\Delta,J_{\Delta,1})+K_1\rho_M^{(i-1)}\left(\Delta,\frac{\epsilon}{3}\right)\right].$$

Then, (C.2) follows from combining the above inequality and the probability fact

$$\begin{aligned}&\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\frac{\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)-\mathcal{L}_i(\Delta,\theta)}{\mathcal{L}_i(\Delta,\theta)}\right|\leq\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)\right)\\&\geq(1-\epsilon)\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\frac{\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)-\mathcal{L}_i(\Delta,\theta)}{\mathcal{L}_i(\Delta,\theta)}\right|\leq\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)\middle|\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}},A_i,A_i'\right).\end{aligned}$$

Apparently, by the definition of $\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)$, we obtain $\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)\to 0$ as $\Delta\to 0$.

For the log-likelihood approximation, we note that

$$\left|\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right|=\left|\log\left(1-\frac{\mathcal{L}_i(\Delta,\theta)-\hat{\mathcal{L}}_i^{(J)}(\Delta,\theta)}{\mathcal{L}_i(\Delta,\theta)}\right)\right|.$$

Taking supremum on both sides, we have

$$\sup_{\theta\in\Theta}\left|\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right|\leq\rho_{\ell}^{(i)}(\Delta,\epsilon),\text{ with }\rho_{\ell}^{(i)}(\Delta,\epsilon)=-\log\left(1-\rho_{\mathcal{L}}^{(i)}(\Delta,\epsilon)\right).$$

Then, (C.3) follows from the continuity of the log and absolute value functions, the convergence of the likelihood update approximation (C.2), and the probabilistic fact

$$\begin{aligned}&\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right|\leq\rho_{\ell}^{(i)}(\Delta,\epsilon)\right)\\&\geq(1-\epsilon)\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right|\leq\rho_{\ell}^{(i)}(\Delta,\epsilon)\middle|\inf_{\theta\in\Theta}\mathcal{L}_i(\Delta,\theta)\geq C_{\mathcal{L}}^{\epsilon/3}\Delta^{-\frac{1}{2}},A_i,A_i'\right).\end{aligned}$$

Apparently, by the definition of $\rho_{\ell}^{(i)}(\Delta,\epsilon)$, we obtain $\rho_{\ell}^{(i)}(\Delta,\epsilon)\to 0$ as $\Delta\to 0$.

For the log-likelihood approximation, we have by definition that

$$\left|\hat{\ell}^{(J)}(\Delta,\theta)-\ell(\Delta,\theta)\right|=\left|\sum_{i=1}^{n}\left(\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right)\right|\leq\sum_{i=1}^{n}\left|\hat{\ell}_i^{(J)}(\Delta,\theta)-\ell_i(\Delta,\theta)\right|.$$

Taking supremum on both sides, we have

$$\sup_{\theta\in\Theta}\left|\hat{\ell}^{(J)}(\Delta,\theta)-\ell(\Delta,\theta)\right|\leq\rho^{(n)}(\Delta,\epsilon),\text{ with }\rho^{(n)}(\Delta,\epsilon)=\sum_{i=1}^{n}\rho_{\ell}^{(i)}(\Delta,\epsilon).$$

Then, (C.4) follows from the convergence of the log-likelihood update approximation (C.3), and the probabilistic fact

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\hat{\ell}^{(J)}(\Delta,\theta)-\ell(\Delta,\theta)\right|\leq\rho^{(n)}(\Delta,\epsilon)\right)$$

$$\geq (1-\epsilon)^n \mathbb{P}\left(\sup_{\theta\in\Theta} \left| \hat{\ell}^{(J)}(\Delta,\theta) - \ell(\Delta,\theta) \right| \leq \rho^{(n)}(\Delta,\epsilon) \middle| \bigcap_{i=1}^{n} \left(\inf_{\theta\in\Theta} \mathcal{L}_i(\Delta,\theta) \geq C_{\mathcal{L}}^{\epsilon/3} \Delta^{-\frac{1}{2}} \right), \bigcap_{i=1}^{n} A_i, \bigcap_{i=1}^{n} A_i' \right).$$

Apparently, by the definition of $\rho^{(n)}(\Delta,\epsilon)$, we obtain $\rho^{(n)}(\Delta,\epsilon) \to 0$ as $\Delta \to 0$.

Step 3 – Convergence of the log-likelihood approximation $\hat{\ell}^{(J)}(\Delta,\theta)$ as $n \to \infty$. For any $\epsilon, \eta > 0$ and fixed n, according to (C.4), there exist a sufficiently small Δ_n^{ϵ}, such that for any $\Delta \leq \Delta_n^{\epsilon}$ and $J \geq J_{\Delta}^{\epsilon}$, we have

$$\mathbb{P}\left(\sup_{\theta\in\Theta} \left| \hat{\ell}^{(J)}(\Delta,\theta) - \ell(\Delta,\theta) \right| \leq \rho^{(n)}(\Delta,\epsilon) \right) \geq 1-\epsilon.$$

Since $\lim_{\Delta\to 0} \rho^{(n)}(\Delta,\epsilon) = 0$, for any $\eta > 0$, there exists a sufficiently small Δ_n^{η}, such that for any $\Delta \leq \Delta_n^{\eta}$, we have $\rho^{(n)}(\Delta,\epsilon) \leq \eta$. This suggests

$$\mathbb{P}\left(\sup_{\theta\in\Theta} \left| \hat{\ell}^{(J)}(\Delta,\theta) - \ell(\Delta,\theta) \right| \leq \eta \right) \geq 1-\epsilon,$$

for any $J \geq J_{\Delta}^{\epsilon}$. Now, we choose $\Delta_n^{\epsilon,\eta} = \min\{\Delta_n^{\epsilon}, \Delta_n^{\eta}, \Delta_{n-1}^{\epsilon,\eta}/2\}$ and $J_n^{\epsilon,\eta} = \max\{J^{\epsilon}_{\Delta_n^{\epsilon,\eta}}, J^{\epsilon}_{\Delta_{n-1}^{\epsilon,\eta}} + 1\}$. By construction, $\Delta_n^{\epsilon,\eta} \to 0$ and $J_n^{\epsilon,\eta} \to \infty$ as $n \to \infty$, and (C.5) holds. □

Appendix C.2.2 A smoothing procedure illustrated under the model of Heston (1993)

When one of Assumptions 1–3 does not hold, we employ the following standard techniques of smoothing to validate our approximation. We will explain why our approximations converge to the corresponding true versions, even if the latent state space $\mathcal{Y}$ is not compact, the derivatives of coefficient functions are not uniformly bounded, or the uniform ellipticity condition does not hold. For illustration purposes, we take the model of Heston (1993) as an example, and explain why our likelihood approximation converges.

The Heston model is specified as

$$dX_t = \left(\mu - \frac{1}{2} Y_t \right) dt + \sqrt{Y_t} dW_{1t}, \tag{C.21a}$$

$$dY_t = \kappa(\alpha - Y_t)dt + \xi\sqrt{Y_t}[\rho dW_{1t} + \sqrt{1-\rho^2} dW_{2t}]. \tag{C.21b}$$

We assume Feller's condition holds for the latent process Y_t, i.e., $2\kappa\alpha \geq \xi^2$. Under the model (C.21a)–(C.21b), all of Assumptions 1–3 are violated: The latent state space $\mathcal{Y} = (0,\infty)$ is not compact; the coefficient function $\sqrt{y}$ is singular at $y = 0$; the uniform ellipticity condition does not hold at $y = 0$.

To validate our approximation, we consider the following limited version of the model with a truncated upper bound U and truncated lower bound L for the latent process Y_t :

$$d\hat{X}_t = \left(\mu - \frac{1}{2} \hat{Y}_t \right) dt + \varphi(\hat{Y}_t) dW_{1t}, \tag{C.22a}$$

$$d\hat{Y}_t = \kappa(\alpha - \hat{Y}_t)dt + \xi\varphi(\hat{Y}_t)[\rho dW_{1t} + \sqrt{1-\rho^2} dW_{2t}], \tag{C.22b}$$

where we assume $\hat{X}_0 = X_0$, $\hat{Y}_0 = Y_0$; the coefficient function $\varphi(\cdot)$ is defined as

$$\varphi(y) = \sqrt{L}1_{\{y \leq L\}} + \sqrt{y}1_{\{L<y\leq U\}} + \sqrt{U}1_{\{y>U\}}.$$

Note that $\varphi(y)$ is not differentiable at $y = L$ and $y = U$. We now smooth the limited model (C.22a)–(C.22b) in what follows, so that Assumptions 1–3 can be satisfied:

$$d\hat{X}_t^\varepsilon = \left(\mu - \frac{1}{2}\hat{Y}_t^\varepsilon\right)dt + \varphi^\varepsilon(\hat{Y}_t^\varepsilon)dW_{1t}, \tag{C.23a}$$

$$d\hat{Y}_t^\varepsilon = \kappa(\alpha - \hat{Y}_t^\varepsilon)dt + \xi\varphi^\varepsilon(\hat{Y}_t^\varepsilon)[\rho dW_{1t} + \sqrt{1-\rho^2}dW_{2t}], \tag{C.23b}$$

with $\hat{X}_0^\varepsilon = X_0$ and $\hat{Y}_0^\varepsilon = Y_0$, where the function $\varphi^\varepsilon(\cdot)$ is a infinitely smooth modification of $\varphi(\cdot)$ explicitly given by

$$\varphi^\varepsilon(y) := \begin{cases} \sqrt{U}, & \text{if } y > U + \varepsilon, \\ f(y-U)\sqrt{y} + (1 - f(y-U))\sqrt{U}, & \text{if } U - \varepsilon < y \leq U + \varepsilon, \\ \sqrt{y}, & \text{if } L + \varepsilon < y \leq U - \varepsilon, \\ (1 - f(y-L))\sqrt{y} + f(y-L)\sqrt{L}, & \text{if } L - \varepsilon < y \leq L + \varepsilon, \\ \sqrt{L}, & \text{if } y \leq L - \varepsilon. \end{cases}$$

Here, the function $f(\cdot)$ is defined as

$$f(x) := \frac{\psi(\varepsilon - x)}{\psi(\varepsilon + x) + \psi(\varepsilon - x)},$$

with $\psi(x) := \exp(-1/x)$ for $x > 0$ and $\psi(x) := 0$ for $x \leq 0$. It is straightforward to verify that, for the smoothed model (C.23a)–(C.23b), the latent state space $[L, U]$ is compact; the derivative of any the coefficient function is uniformly bounded with respect to x, y, and θ; the uniform ellipticity condition (B.1a)–(B.1b) holds with $a_1 = L$ and $a_2 = U$.

According to Theorem 2, under the smoothed model (C.23a)–(C.23b), our likelihood approximation converges to the true likelihood in probability. Now, it suffices to show the likelihood function induced by the smoothed model (C.23a)–(C.23b) converges to that induced by the true model (C.21a)–(C.21b) in probability as well.[12] More precisely, we will show

$$p^\varepsilon(\mathbf{X}_{n\Delta}) \xrightarrow{p} p(\mathbf{X}_{n\Delta}), \text{ as } L \to 0, \ U \to \infty, \text{ and } \varepsilon \to 0,$$

[12] In practice, one does not need to smooth the Heston model before implementing our approximations. Our implementation under the original Heston model following the algorithm provided in this paper can be viewed as an implementation under a smoothed Heston model. Indeed, according to the adaptive algorithm for choosing the grid points introduced in Appendix A.2, one needs to truncate the latent space $\mathcal{Y}$ as $[y^{(0)}, y^{(m)}]$ before determining the basis functions. By setting $y^{(0)} = L + \varepsilon$ and $y^{(m)} = U - \varepsilon$, all the intermediate components (Euler approximations of the marginal transition density p_X and truncated marginal transition moments $B_{(y^{(k)},y^{(k+1)}),l}$, as well as the induced interpolation coefficients $\alpha_{(y^{(k)},y^{(k+1)}),l}$ and $\beta^{(y^{(j)},y^{(j+1)}),\ell}_{(y^{(k)},y^{(k+1)}),l}$) are the same under the smoothed Heston model and the true Heston model with the space truncation. Thus, the resulting approximations of the likelihood (resp. filters) under these two models coincide with each other.

where $p^{\varepsilon}(\mathbf{X}_{n\Delta})$ (resp. $p(\mathbf{X}_{n\Delta})$) represents the likelihood under the smoothed model (C.23a)–(C.23b) (resp. true Heston model (C.21a)–(C.21b).)

For this, we introduce a stopping time

$$\tau := \inf \left\{ t \geq 0 : Y_t \leq L + \varepsilon \text{ or } Y_t \geq U - \varepsilon \right\}.$$

By the construction of the smoothed model (C.23a)–(C.23b), we further have

$$\tau \equiv \inf \left\{ t \geq 0 : Y_t^{\varepsilon} \leq L + \varepsilon \text{ or } Y_t^{\varepsilon} \geq U - \varepsilon \right\},$$

which immediately implies $\mathbf{X}_{n\Delta}^{\varepsilon} = \mathbf{X}_{n\Delta}$ for $n\Delta < \tau$. Denote by $p^{\varepsilon}(\mathbf{x}_{n\Delta})$ (resp. $p(\mathbf{x}_{n\Delta})$) the joint density of $\mathbf{X}_{n\Delta}^{\varepsilon}$ (resp. $\mathbf{X}_{n\Delta}$), under the smoothed model (C.23a)–(C.23b) (resp. true Heston model (C.21a)–(C.21b)). Then, conditioning on $n\Delta < \tau$, we have

$$p^{\varepsilon}(\mathbf{x}_{n\Delta} | n\Delta < \tau) = p(\mathbf{x}_{n\Delta} | n\Delta < \tau), \text{ for any } \mathbf{x}_{n\Delta} \in \mathcal{X}^n. \tag{C.24}$$

For any $\eta > 0$, we note that

$$\begin{aligned}
&\mathbb{P}\left(|p^{\varepsilon}(\mathbf{X}_{n\Delta}) - p(\mathbf{X}_{n\Delta})| \geq \eta\right) \\
&= \mathbb{P}\left(|p^{\varepsilon}(\mathbf{X}_{n\Delta}) - p(\mathbf{X}_{n\Delta})| \geq \eta, n\Delta < \tau\right) + \mathbb{P}\left(|p^{\varepsilon}(\mathbf{X}_{n\Delta}) - p(\mathbf{X}_{n\Delta})| \geq \eta, n\Delta \geq \tau\right) \\
&\leq \mathbb{P}\left(|p^{\varepsilon}(\mathbf{X}_{n\Delta}) - p(\mathbf{X}_{n\Delta})| \geq \eta | n\Delta < \tau\right) \mathbb{P}\left(n\Delta < \tau\right) + \mathbb{P}\left(n\Delta \geq \tau\right).
\end{aligned}$$

According to (C.24), we have

$$\mathbb{P}\left(|p^{\varepsilon}(\mathbf{X}_{n\Delta}) - p(\mathbf{X}_{n\Delta})| \geq \eta | n\Delta < \tau\right) = 0,$$

which implies

$$\mathbb{P}\left(|p^{\varepsilon}(\mathbf{X}_{n\Delta}) - p(\mathbf{X}_{n\Delta})| \geq \eta\right) \leq \mathbb{P}\left(n\Delta \geq \tau\right).$$

Now, it suffices to show $\mathbb{P}\left(n\Delta \geq \tau\right) \to 0$ as $L \to 0$, $U \to \infty$, and $\varepsilon \to 0$.

Note that

$$\mathbb{P}\left(n\Delta \geq \tau\right) \leq \mathbb{P}\left(\max_{0 \leq t \leq n\Delta} Y_t \geq U - \varepsilon\right) + \mathbb{P}\left(\min_{0 \leq t \leq n\Delta} Y_t \leq L + \varepsilon\right).$$

Thanks to the Feller condition $2\kappa\alpha \geq \xi^2$, the CIR process Y_t cannot attain zero or positive infinity with probability 1 (see, e.g., Section 5.5 of Karatzas and Shreve (1991)). Thus, we have

$$\mathbb{P}\left(\max_{0 \leq t \leq n\Delta} Y_t \geq U - \varepsilon\right) \to 0, \text{ as } U \to \infty, \text{ and } \mathbb{P}\left(\min_{0 \leq t \leq n\Delta} Y_t \leq L + \varepsilon\right) \to 0, \text{ as } L \to 0 \text{ and } \varepsilon \to 0.$$

which implies

$$\mathbb{P}\left(n\Delta \geq \tau\right) \to 0, \text{ as } L \to 0, \ U \to \infty, \text{ and } \varepsilon \to 0.$$

Panel A: Comparison of AMMLE $\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}}$ and true values θ_0

		1000 obs	2500 obs	5000 obs	10000 obs	20000 obs
Parameter	True	Bias	Bias	Bias	Bias	Bias
		(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)
		[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]
μ	0.05	-0.0067	7.0×10^{-5}	0.0065	0.0049	0.0038
		(0.152)	(0.0967)	(0.0690)	(0.0491)	(0.0329)
		[0.149]	[0.0964]	[0.0682]	[0.0483]	[0.0342]
κ	3	1.5	0.40	0.19	0.067	0.016
		(5.82)	(1.42)	(0.868)	(0.569)	(0.394)
		[2.478]	[1.200]	[0.789]	[0.526]	[0.359]
α	0.1	5.7×10^{-4}	3.0×10^{-4}	-3.4×10^{-4}	-2.6×10^{-4}	-2.6×10^{-4}
		(0.0139)	(0.00886)	(0.00614)	(0.00433)	(0.00296)
		[0.0126]	[0.00822]	[0.00583]	[0.00418]	[0.00298]
ξ	0.25	-6.2×10^{-4}	-0.0013	-0.0039	-0.0054	-0.0055
		(0.0981)	(0.0560)	(0.0382)	(0.0261)	(0.0177)
		[0.0871]	[0.0507]	[0.0351]	[0.0244]	[0.0171]
ρ	-0.7	-0.073	-0.017	-6.1×10^{-4}	0.0044	0.0089
		(0.229)	(0.123)	(0.0829)	(0.0571)	(0.0377)
		[0.180]	[0.108]	[0.0766]	[0.0540]	[0.0381]

Panel B: Comparison of AFMLE $\hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}$ and true values θ_0

		1000 obs	2500 obs	5000 obs	10000 obs	20000 obs
Parameter	True	Bias	Bias	Bias	Bias	Bias
		(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)
		[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]
μ	0.05	-0.012	-0.0033	0.0021	5.9×10^{-4}	-3.0×10^{-4}
		(0.151)	(0.0958)	(0.0692)	(0.0485)	(0.0323)
		[0.152]	[0.0956]	[0.0674]	[0.0476]	[0.0336]
κ	3	0.58	0.19	0.11	0.0054	0.027
		(1.12)	(0.617)	(0.420)	(0.291)	(0.209)
		[1.02]	[0.590]	[0.410]	[0.284]	[0.199]
α	0.1	6.1×10^{-4}	5.7×10^{-4}	9.3×10^{-6}	-1.9×10^{-5}	-2.8×10^{-5}
		(0.0136)	(0.00819)	(0.00568)	(0.00397)	(0.00271)
		[0.0122]	[0.00805]	[0.00568]	[0.00408]	[0.00287]
ξ	0.25	3.5×10^{-4}	3.1×10^{-4}	2.7×10^{-4}	1.8×10^{-4}	1.2×10^{-4}
		(0.00513)	(0.00314)	(0.00235)	(0.00158)	(0.00111)
		[0.00493]	[0.00311]	[0.00220]	[0.00155]	[0.00110]
ρ	-0.7	-7.0×10^{-4}	-1.8×10^{-4}	-3.9×10^{-5}	-9.2×10^{-5}	2.4×10^{-5}
		(0.0142)	(0.00928)	(0.00612)	(0.00439)	(0.00323)
		[0.0141]	[0.00889]	[0.00629]	[0.00444]	[0.00314]

Panel C: Comparison of AMMLE $\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}}$ and AFMLE $\hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}$

		1000 obs	2500 obs	5000 obs	10000 obs	20000 obs
Parameter	True	Bias	Bias	Bias	Bias	Bias
		(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)
μ	0.05	0.0050	0.0034	0.0044	0.0043	0.0042
		(0.0538)	(0.0234)	(0.0146)	(0.00964)	(0.00650)
κ	3	0.89	0.21	0.076	0.013	-0.010
		(5.65)	(1.27)	(0.772)	(0.504)	(0.342)
α	0.1	-4.2×10^{-5}	-2.7×10^{-4}	-3.5×10^{-4}	-2.4×10^{-4}	-2.4×10^{-4}
		(0.00764)	(0.00391)	(0.00251)	(0.00168)	(0.00117)
ξ	0.25	-1.4×10^{-4}	-0.0016	-0.0042	-0.0056	-0.0056
		(0.0973)	(0.0557)	(0.0380)	(0.0259)	(0.0175)
ρ	-0.7	-0.073	-0.017	-5.8×10^{-4}	0.0045	0.0088
		(0.229)	(0.122)	(0.0826)	(0.0572)	(0.0375)

Table 1: Monte Carlo results for the Heston model at daily frequency

Note: For all these three panels, the header "F. std. dev." (resp. "A. std. dev.") is the abbreviation of the finite-sample standard deviation (resp. asymptotic standard deviation). In Panels A (resp. B), each bias and finite-sample standard deviation is computed based on 500 estimators, while the asymptotic standard deviation averages 500 sample-based asymptotic standard deviations calculated according to (42) (resp. (43)). In Panel C, each bias (resp. finite-sample standard deviation) is computed as the mean (resp. standard deviation) of 500 estimators of $\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}} - \hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}$.

Panel A: Comparison of AMMLE $\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}}$ and true values θ_0

		1000 obs	2500 obs	5000 obs	10000 obs	20000 obs
Parameter	True	Bias	Bias	Bias	Bias	Bias
		(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)
		[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]
μ	0.05	0.0013	0.0050	0.0077	0.0062	0.0056
		(0.0698)	(0.0442)	(0.0316)	(0.0224)	(0.0151)
		[0.0706]	[0.0446]	[0.0315]	[0.0223]	[0.0158]
κ	3	0.49	0.087	-0.011	-0.027	-0.052
		(1.97)	(0.904)	(0.576)	(0.410)	(0.283)
		[1.524]	[0.840]	[0.570]	[0.393]	[0.274]
α	0.1	-2.3×10^{-5}	-5.3×10^{-5}	-3.5×10^{-4}	-2.1×10^{-4}	-2.0×10^{-4}
		(0.00782)	(0.00495)	(0.00343)	(0.00239)	(0.00165)
		[0.00718]	[0.00455]	[0.00321]	[0.00228]	[0.00161]
ξ	0.25	-0.015	-0.012	-0.013	-0.012	-0.012
		(0.0697)	(0.0483)	(0.0338)	(0.0232)	(0.0163)
		[0.0779]	[0.0468]	[0.0328]	[0.0229]	[0.0162]
ρ	-0.7	-0.040	0.0033	0.017	0.023	0.027
		(0.176)	(0.110)	(0.0714)	(0.0489)	(0.0329)
		[0.185]	[0.103]	[0.0707]	[0.0495]	[0.0347]

Panel B: Comparison of AFMLE $\hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}$ and true values θ_0

		1000 obs	2500 obs	5000 obs	10000 obs	20000 obs
Parameter	True	Bias	Bias	Bias	Bias	Bias
		(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)
		[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]	[A. std. dev.]
μ	0.05	-0.0027	7.3×10^{-4}	0.0028	0.0011	4.7×10^{-4}
		(0.0678)	(0.0430)	(0.0311)	(0.0220)	(0.0148)
		[0.0687]	[0.0434]	[0.0306]	[0.0216]	[0.0153]
κ	3	0.11	0.028	0.017	-3.2×10^{-5}	-0.0024
		(0.440)	(0.260)	(0.189)	(0.130)	(0.0963)
		[0.429]	[0.264]	[0.186]	[0.131]	[0.0923]
α	0.1	1.2×10^{-4}	1.3×10^{-4}	-1.0×10^{-4}	-6.7×10^{-5}	-4.8×10^{-5}
		(0.00610)	(0.00366)	(0.00253)	(0.00179)	(0.00125)
		[0.00581]	[0.00372]	[0.00262]	[0.00186]	[0.00131]
ξ	0.25	4.9×10^{-4}	3.9×10^{-4}	3.6×10^{-4}	2.6×10^{-4}	2.1×10^{-4}
		(0.00516)	(0.00320)	(0.00238)	(0.00160)	(0.00111)
		[0.00501]	[0.00316]	[0.00223]	[0.00158]	[0.00111]
ρ	-0.7	-7.2×10^{-4}	-1.4×10^{-4}	5.4×10^{-5}	-6.7×10^{-5}	4.5×10^{-5}
		(0.0142)	(0.00921)	(0.00611)	(0.00441)	(0.00323)
		[0.0141]	[0.00891]	[0.00630]	[0.00444]	[0.00314]

Panel C: Comparison of AMMLE $\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}}$ and AFMLE $\hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}$						
		1000 obs	2500 obs	5000 obs	10000 obs	20000 obs
Parameter	True	Bias	Bias	Bias	Bias	Bias
		(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)	(F. std. dev.)
μ	0.05	0.0040	0.0046	0.0054	0.0053	0.0052
		(0.0171)	(0.0103)	(0.00717)	(0.00491)	(0.00347)
κ	3	0.38	0.048	-0.031	-0.034	-0.054
		(1.90)	(0.881)	(0.570)	(0.389)	(0.265)
α	0.1	-1.1×10^{-4}	-2.1×10^{-4}	-2.6×10^{-4}	-1.7×10^{-4}	-1.6×10^{-4}
		(0.00502)	(0.00315)	(0.00220)	(0.00155)	(0.00108)
ξ	0.25	-0.016	-0.013	-0.013	-0.012	-0.012
		(0.0693)	(0.0484)	(0.0335)	(0.0233)	(0.0159)
ρ	-0.7	-0.040	0.0038	0.018	0.024	0.027
		(0.176)	(0.109)	(0.0712)	(0.0489)	(0.0327)

Table 2: Monte Carlo results for the Heston model at weekly frequency

Note: Except for changing the observation frequency to weekly, all the other setting and methods of calculation for producing these three panels are the same as those for producing Table 1.

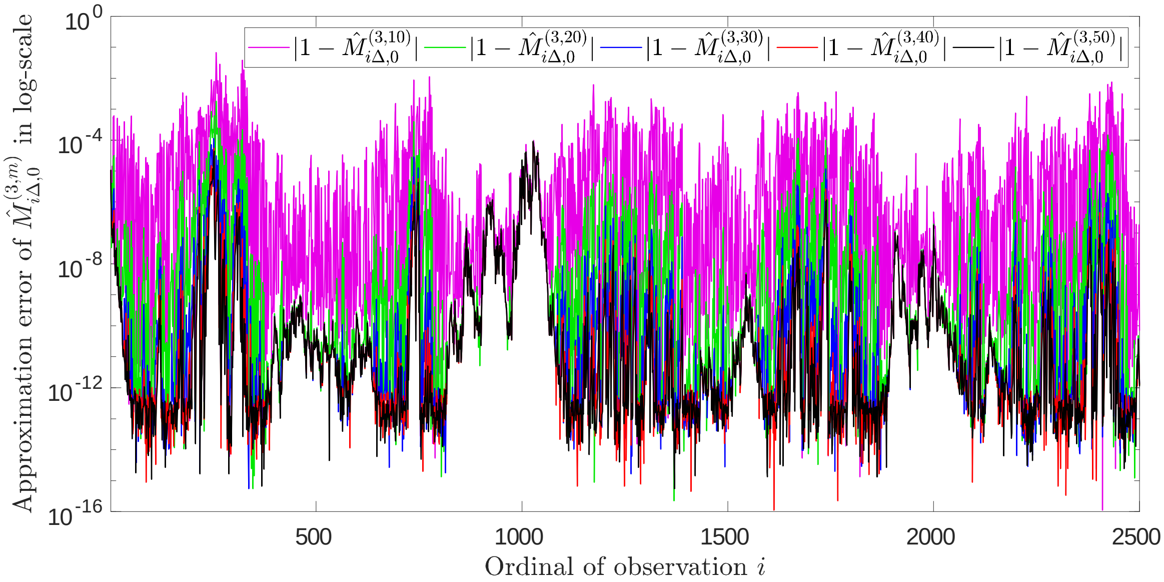

Figure 1: The Heston model: Convergence of the zeroth order filtered moment approximations

Note: All the approximations $\hat{M}^{(3,m)}_{i\Delta,0}$ of zeroth order filtered moment are calculated according to (41a) and (34) with $l = 0$ under the piecewise monomial basis functions $b_{(y^{(k)},y^{(k+1)}),\ell}$ for $\ell = 0, 1, 2, 3$ and $k = 0, 1, 2, \ldots, m$. Here, the grid points $\{y^{(k)}\}_{k=0}^{m}$ are equidistant in $(0.01, 0.3]$ with $y^{(0)} = 0.01$ and $y^{(m)} = 0.3$.

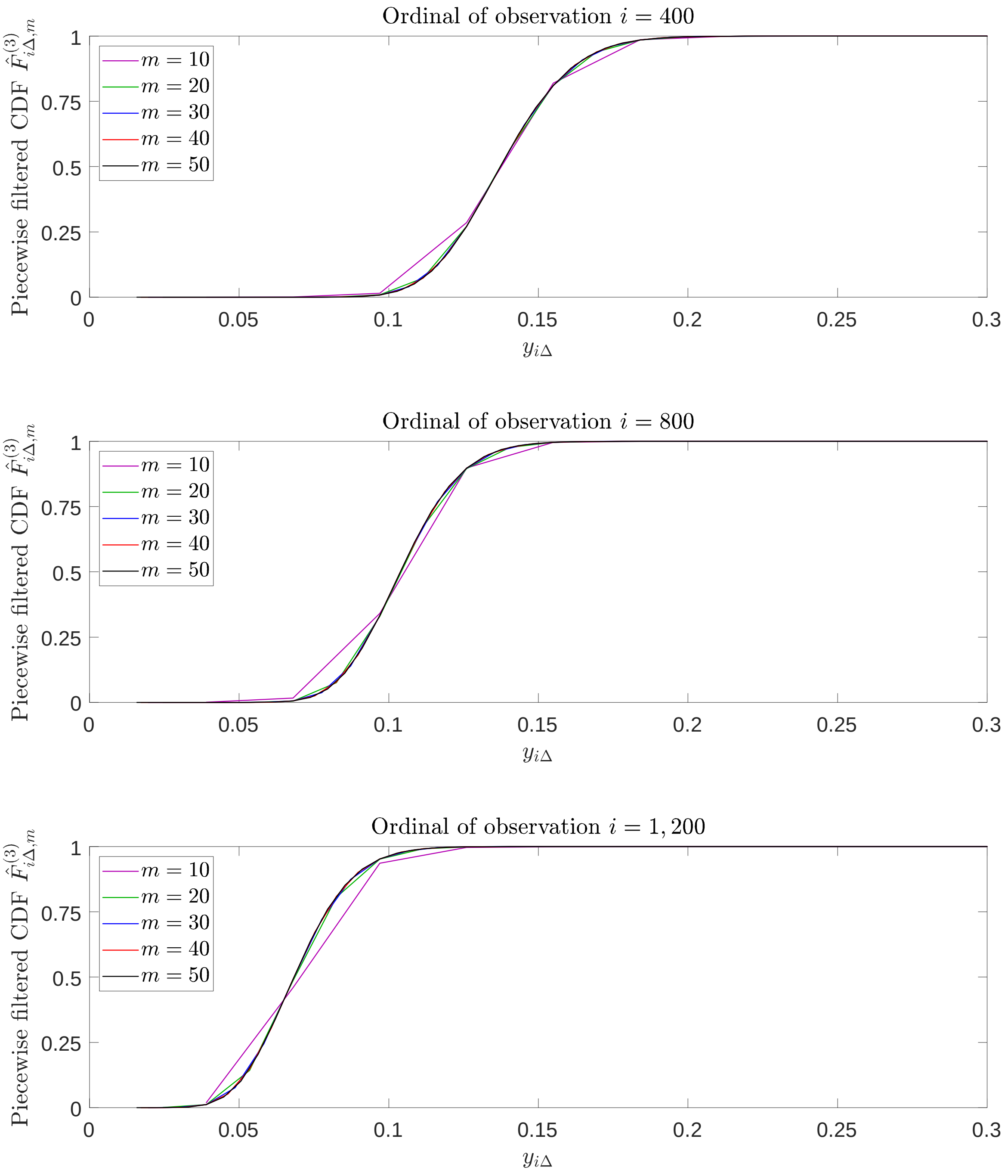

Figure 2: The Heston model: Convergence of piecewise filtered CDF approximations on various dates (I)

Note: Except for corresponding to the approximations of piecewise filtered CDF $\hat{F}^{(3)}_{i\Delta,m}$ given by (41a) and (32) for $i = 400$, 800, and $1,200$, all the other settings for producing these three panels are the same as those for producing Figure 1.

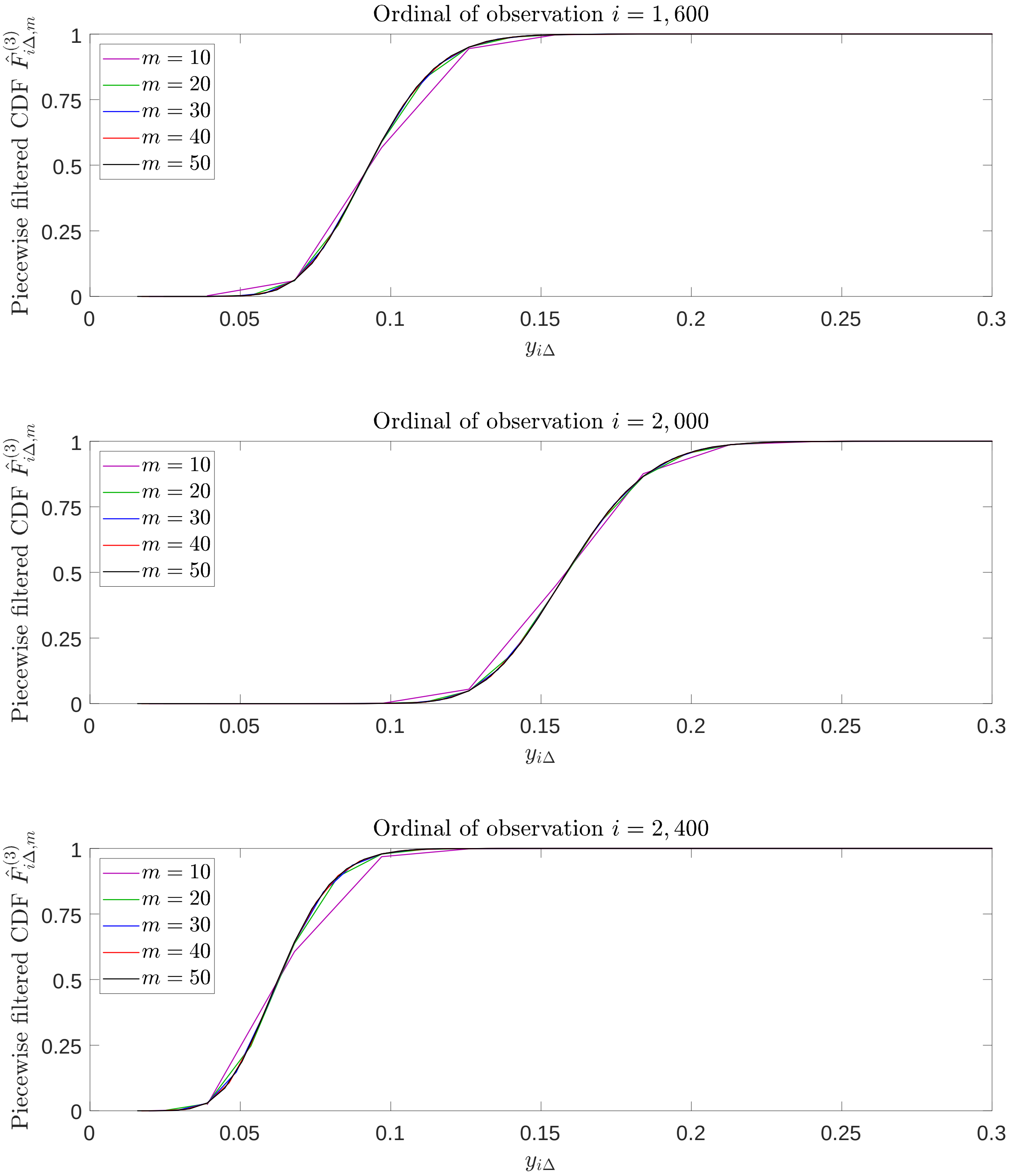

Figure 3: The Heston model: Convergence of piecewise filtered CDF approximations on various dates (II)

Note: Except for corresponding to the cases with $i = 1,600$, $2,000$, and $2,400$, all the other settings for producing these three panels are the same as those for producing Figure 2.

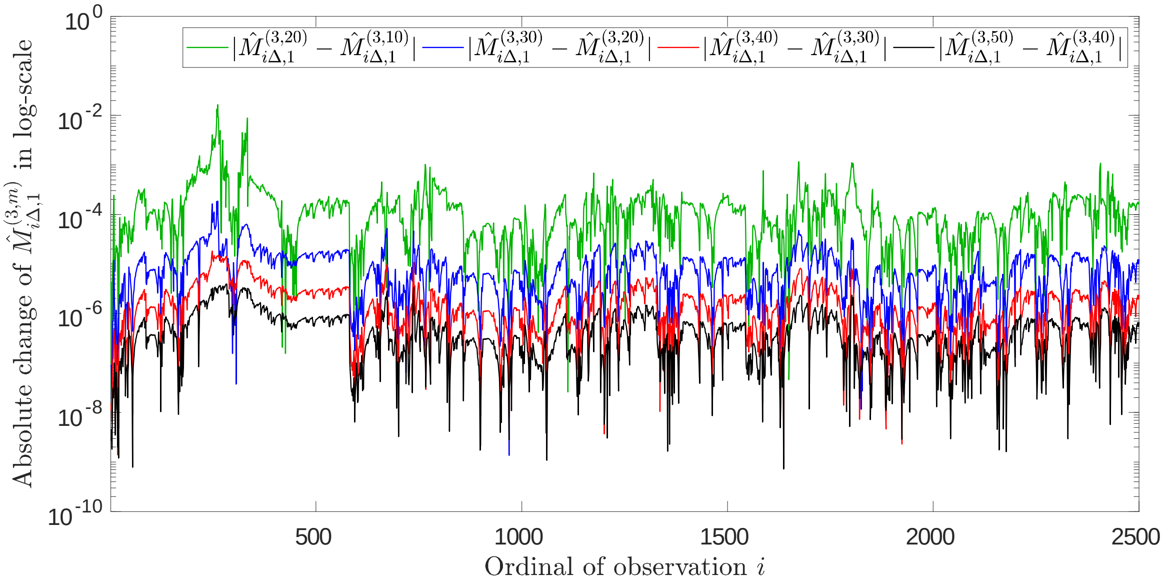

Figure 4: Convergence of the first order filtered moment approximations

Note: Except for corresponding to the approximations of first order filtered moment $\hat{M}_{i\Delta,1}^{(3,m)}$ given by (41a) and (34) with $l = 1$, all the other settings for producing this figure are the same as those for producing Figure 1.

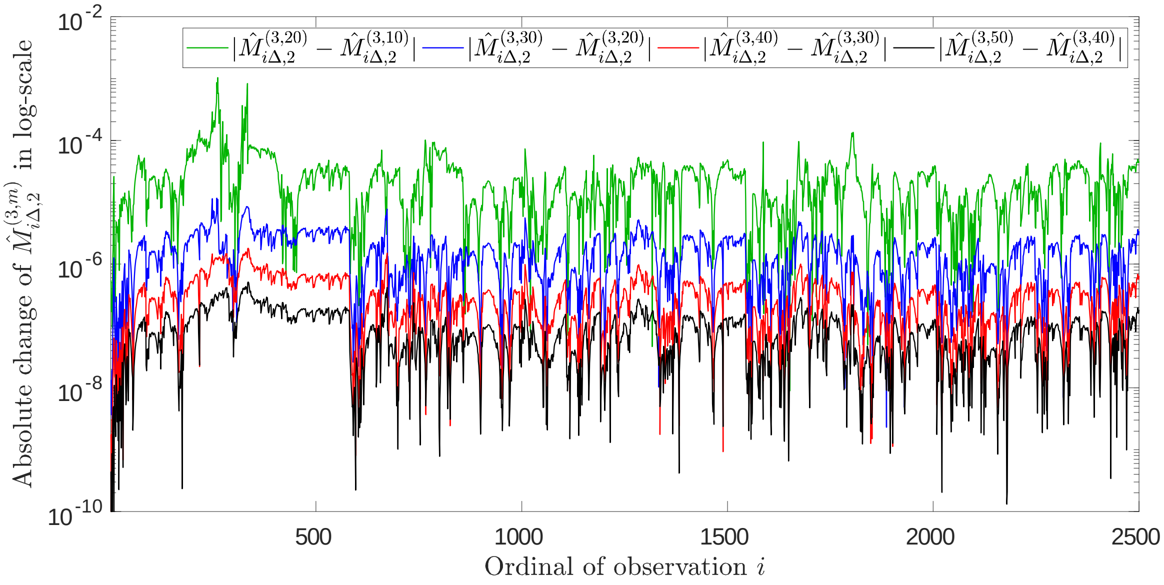

Figure 5: The Heston model: Convergence of the second order filtered moment approximations

Note: Except for corresponding to the approximations of second order filtered moment $\hat{M}_{i\Delta,2}^{(3,m)}$ given by (41a) and (34) with $l = 2$, all the other settings for producing this figure are the same as those for producing Figure 1.

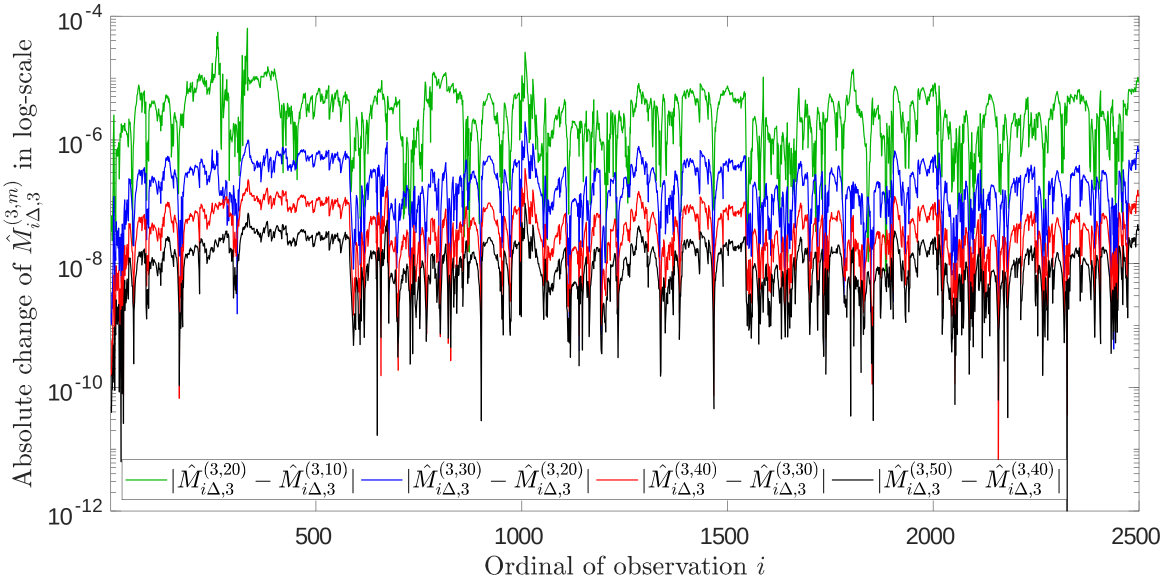

Figure 6: The Heston model: Convergence of the third order filtered moment approximations

Note: Except for corresponding to the approximations of third order filtered moment $\hat{M}^{(3,m)}_{i\Delta,3}$ given by (41a) and (34) with $l = 3$, all the other settings for producing this figure are the same as those for producing Figure 1.

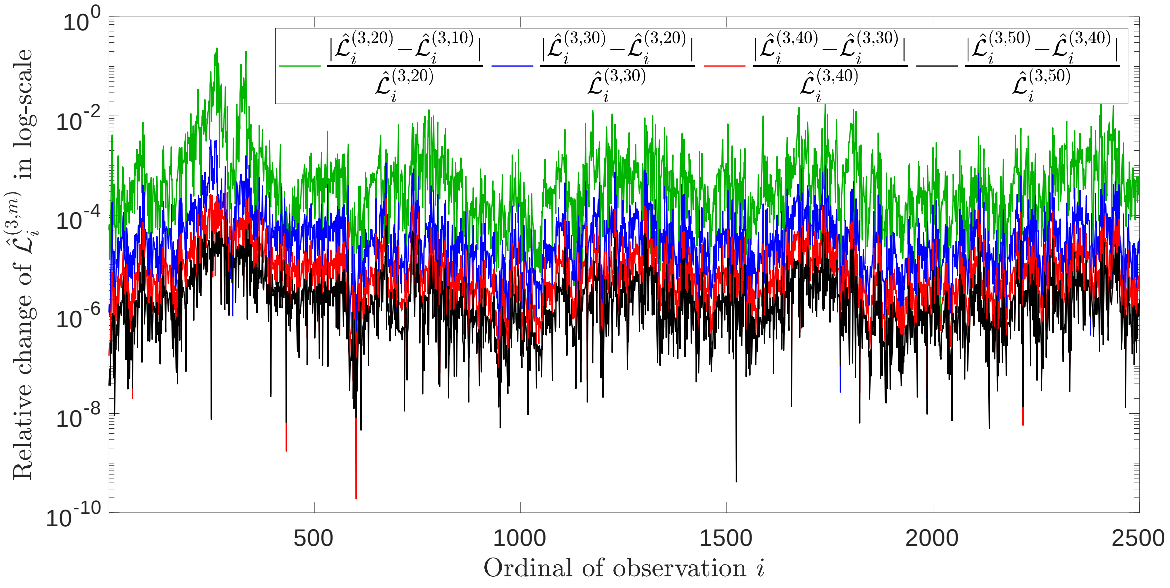

Figure 7: The Heston model: Convergence of the likelihood update approximations

Note: Except for corresponding to the approximations of likelihood update $\hat{\mathcal{L}}^{(3,m)}_i$ given by (41a) and (29), all the other settings for producing this figure are the same as those for producing Figure 1.

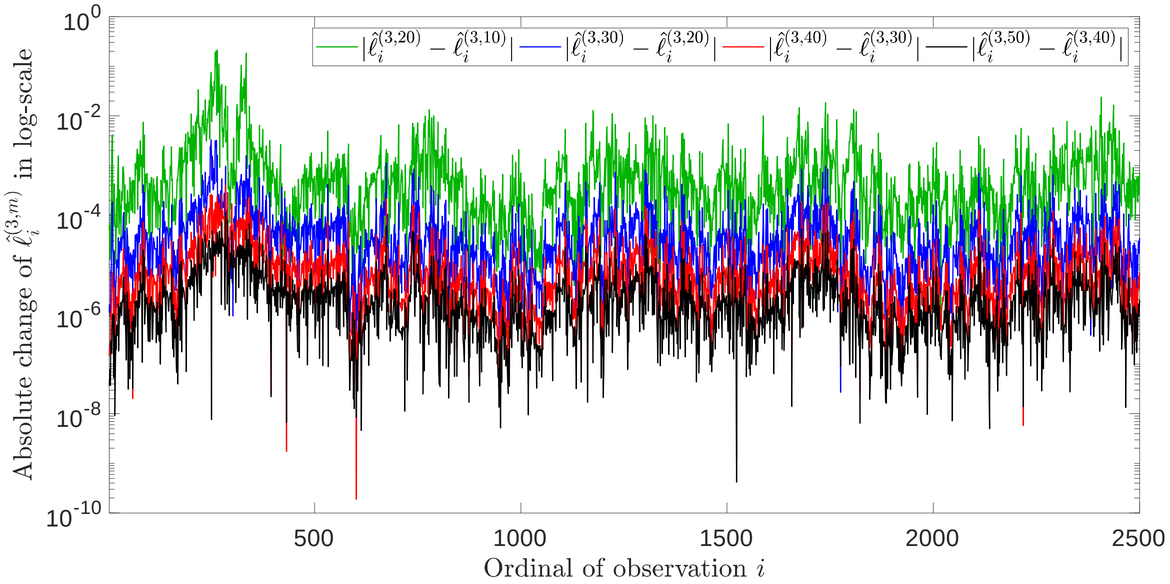

Figure 8: The Heston model: Convergence of the log-likelihood update approximations

Note: Except for corresponding to the approximations of log-likelihood update $\hat{\ell}_i^{(3,m)}$ given by (41b) and (29), all the other settings for producing this figure are the same as those for producing Figure 1.

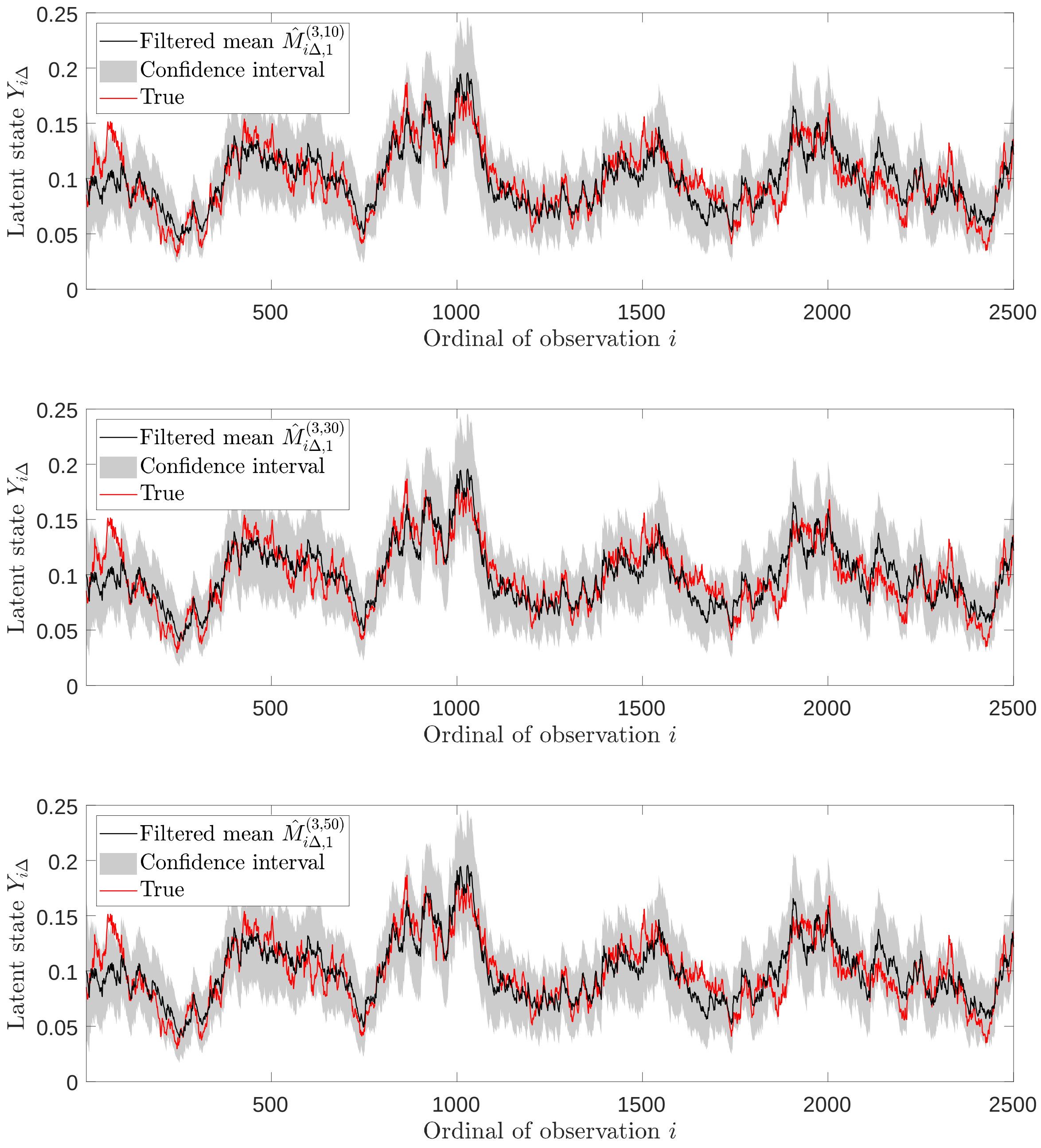

Figure 9: Comparison: filtered versus true latent states with various orders of approximations under the Heston model

Note: In each panel, the filtered mean $\hat{M}^{(3,m)}_{i\Delta,1}$ in black solid curve is calculated according to (41a) and (34) with $l = 1$. For each i, the confidence interval in shadow is constructed by shifting the filtered mean upward and downward twice the filtered standard deviation $\sqrt{\hat{M}^{(3,m)}_{i\Delta,2} - (\hat{M}^{(3,m)}_{i\Delta,1})^2}$. The curve in red represents the true simulated values of latent states generated simultaneously with observations $\{X_{i\Delta}\}_{i=0}^{n}$.

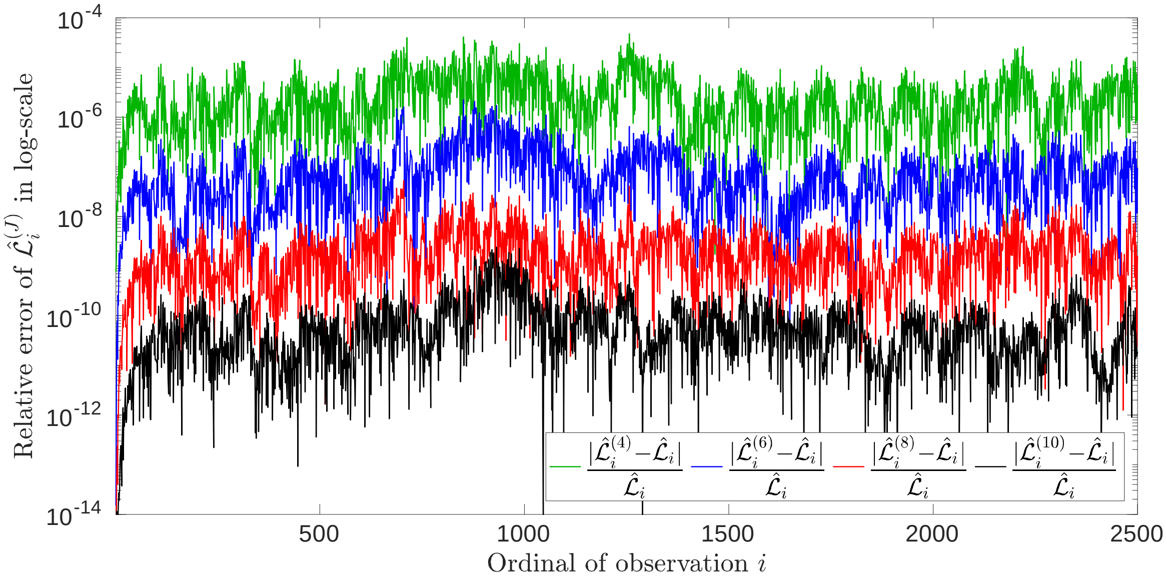

Figure 10: The BOU model: Convergence of the likelihood update approximations

Note: The benchmark of $\mathcal{L}_i$ is computed by the exact Kalman filter, while the approximation $\hat{\mathcal{L}}_i^{(J)}$ is computed according to (14) based on the monomial basis functions $\{b_k\}_{k=0}^J$ with $b_k(y;\theta) = y^k$.

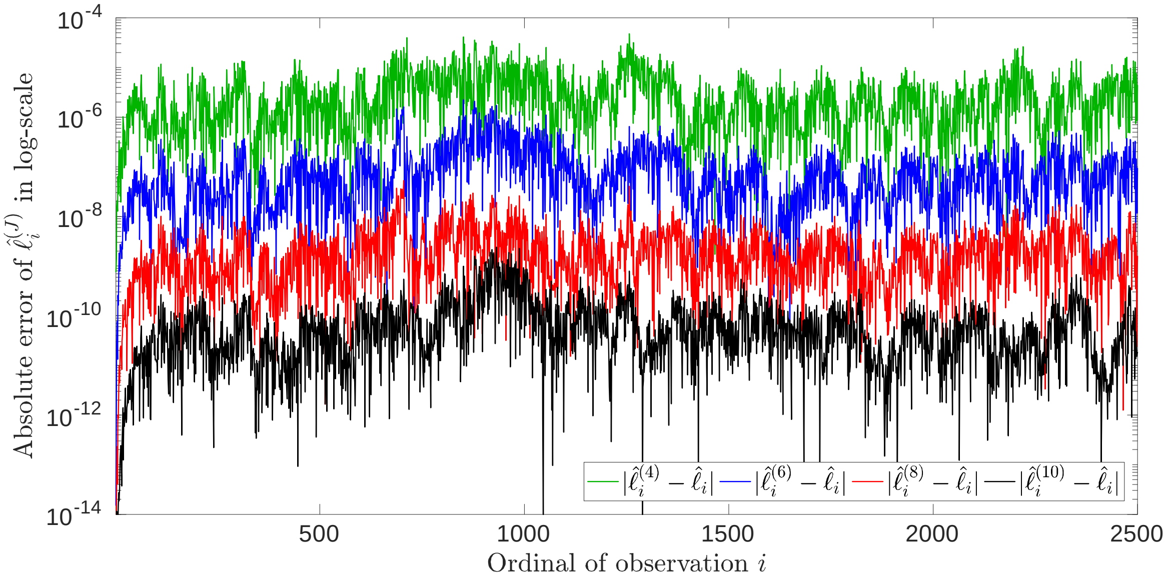

Figure 11: The BOU model: Convergence of the log-likelihood update approximations

Note: Except for corresponding to the approximations of log-likelihood update $\hat{\ell}_i^{(J)}$ given by (17), all the other settings for producing this figure are the same as those for producing Figure 10.

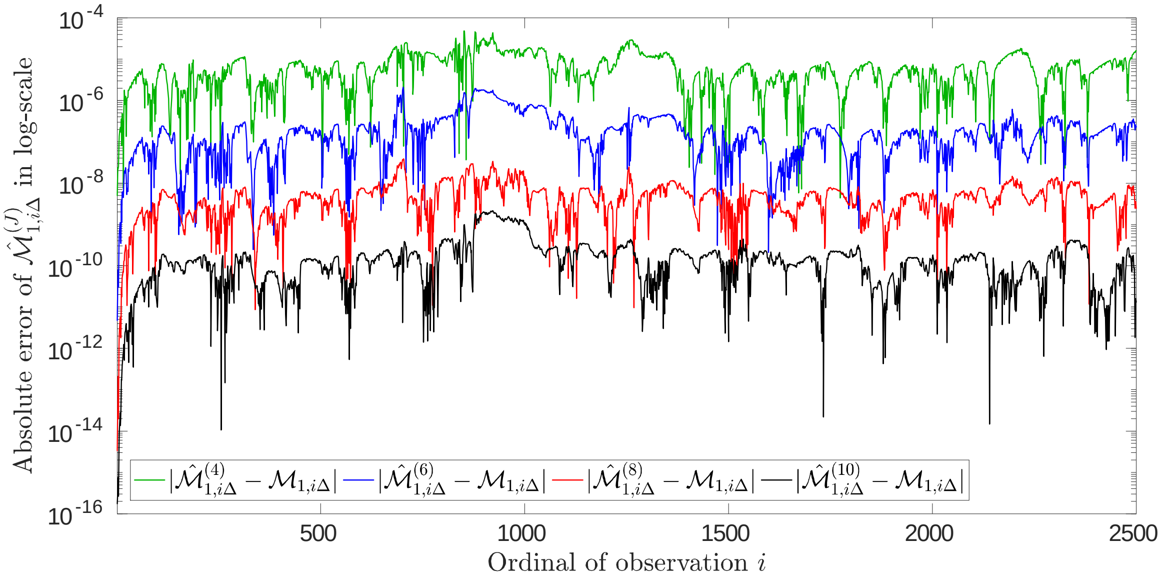

Figure 12: The BOU model: Convergence of the first order filtered moment approximations

Note: Except for corresponding to the approximations of first order filtered moment $\hat{\mathcal{M}}^{(J)}_{1,i\Delta}$ given by (15) with $k = 1$, all the other settings for producing this figure are the same as those for producing Figure 10.

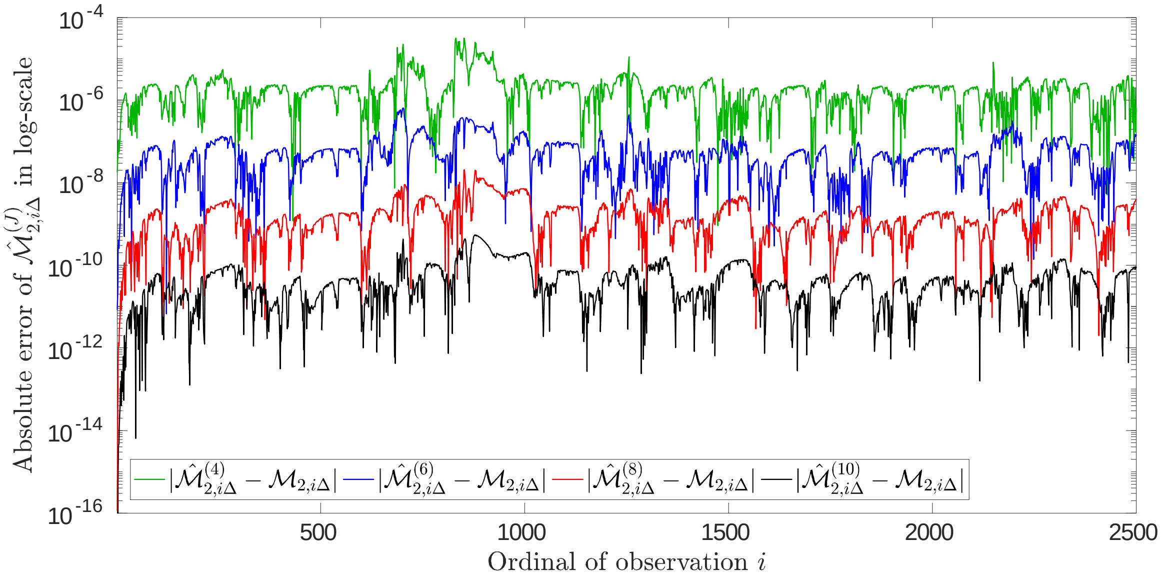

Figure 13: The BOU model: Convergence of the second order filtered moment approximations

Note: Except for corresponding to the approximations of first order filtered moment $\hat{\mathcal{M}}^{(J)}_{2,i\Delta}$ given by (15) with $k = 2$, all the other settings for producing this figure are the same as those for producing Figure 10.

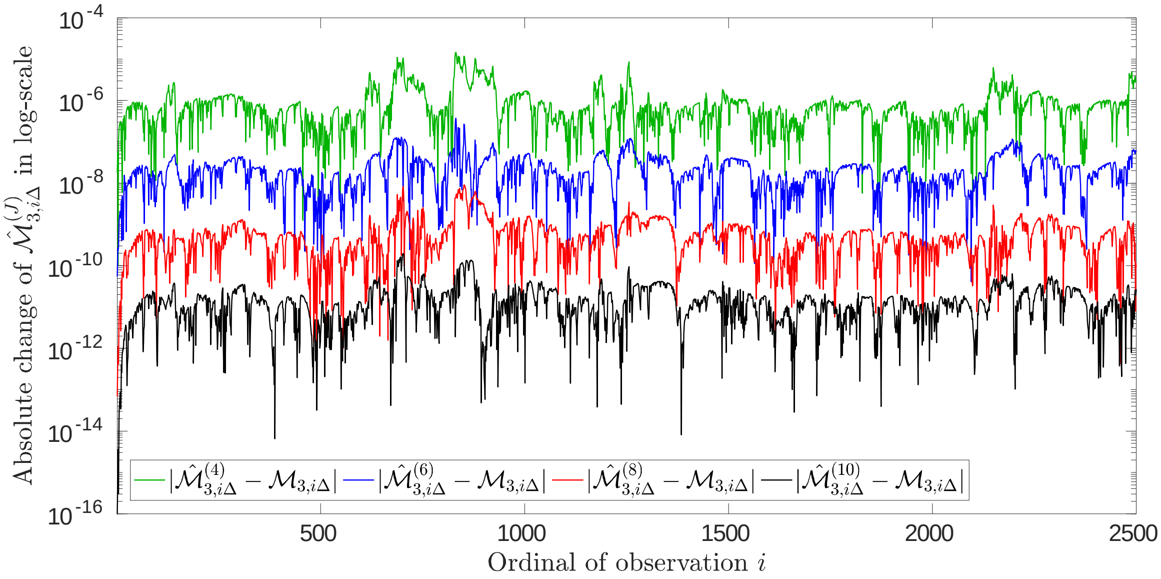

Figure 14: The BOU model: Convergence of the third order filtered moment approximations

Note: Except for corresponding to the approximations of first order filtered moment $\hat{\mathcal{M}}^{(J)}_{3,i\Delta}$ given by (15) with $k = 3$, all the other settings for producing this figure are the same as those for producing Figure 10.

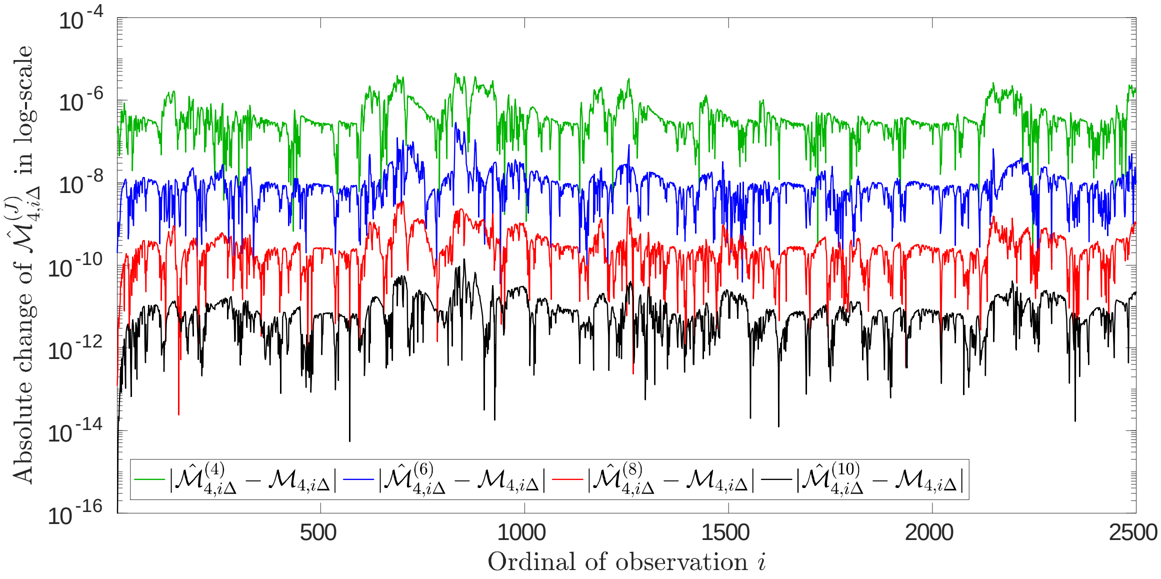

Figure 15: The BOU model: Convergence of the fourth order filtered moment approximations

Note: Except for corresponding to the approximations of first order filtered moment $\hat{\mathcal{M}}^{(J)}_{4,i\Delta}$ given by (15) with $k = 4$, all the other settings for producing this figure are the same as those for producing Figure 10.

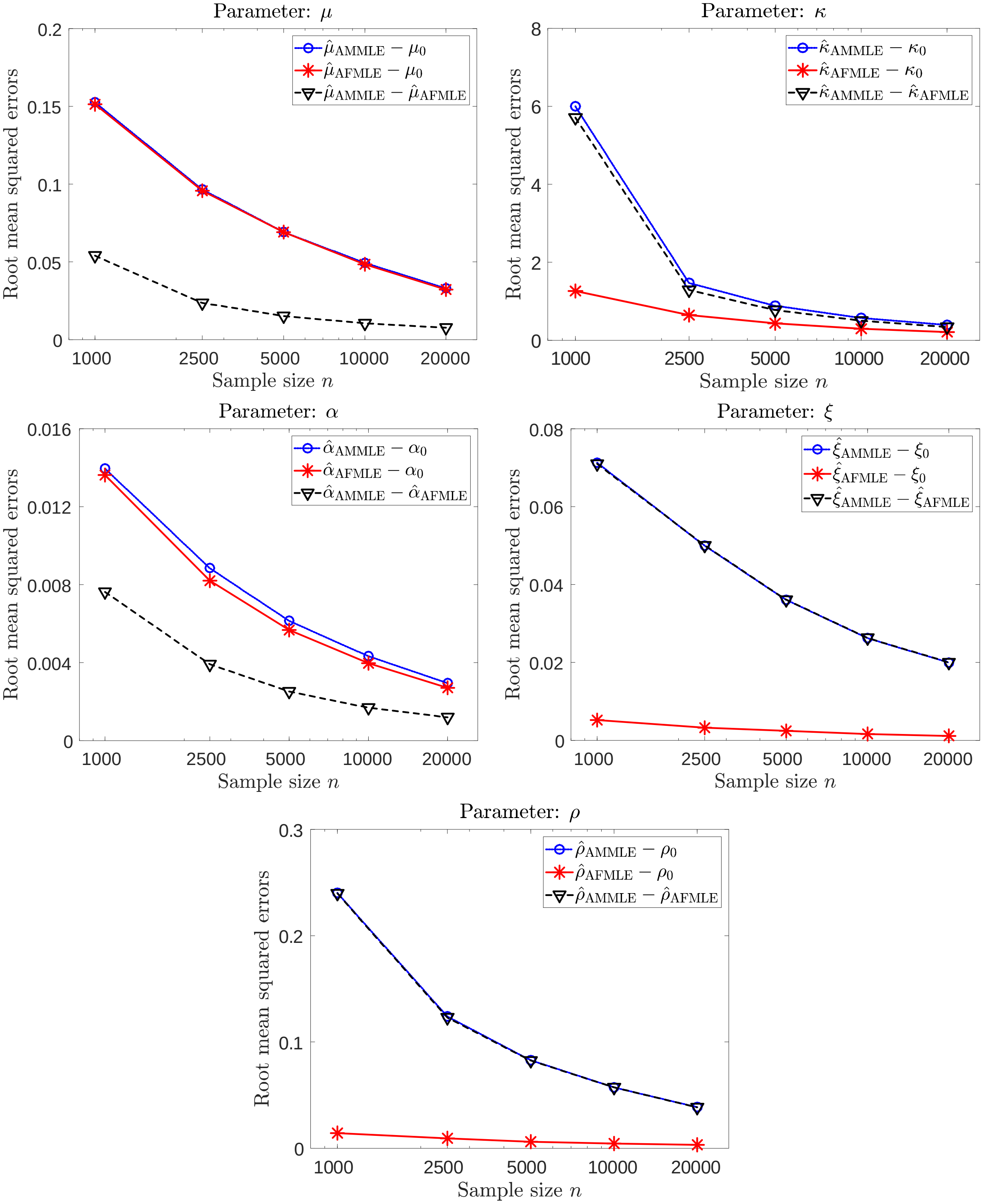

Figure 16: Comparison of AMMLE and AFMLE at daily frequency

Note: In each panel, each RMSE marked by a blue circle (resp. red star) is calculated based on the 500 AMMLEs (resp. AFMLEs), while each RMSE marked by a black triangle is calculated based on the 500 differenced estimators $\hat{\theta}^{(n,\Delta)}_{\text{AMMLE}} - \hat{\theta}^{(n,\Delta)}_{\text{AFMLE}}$.

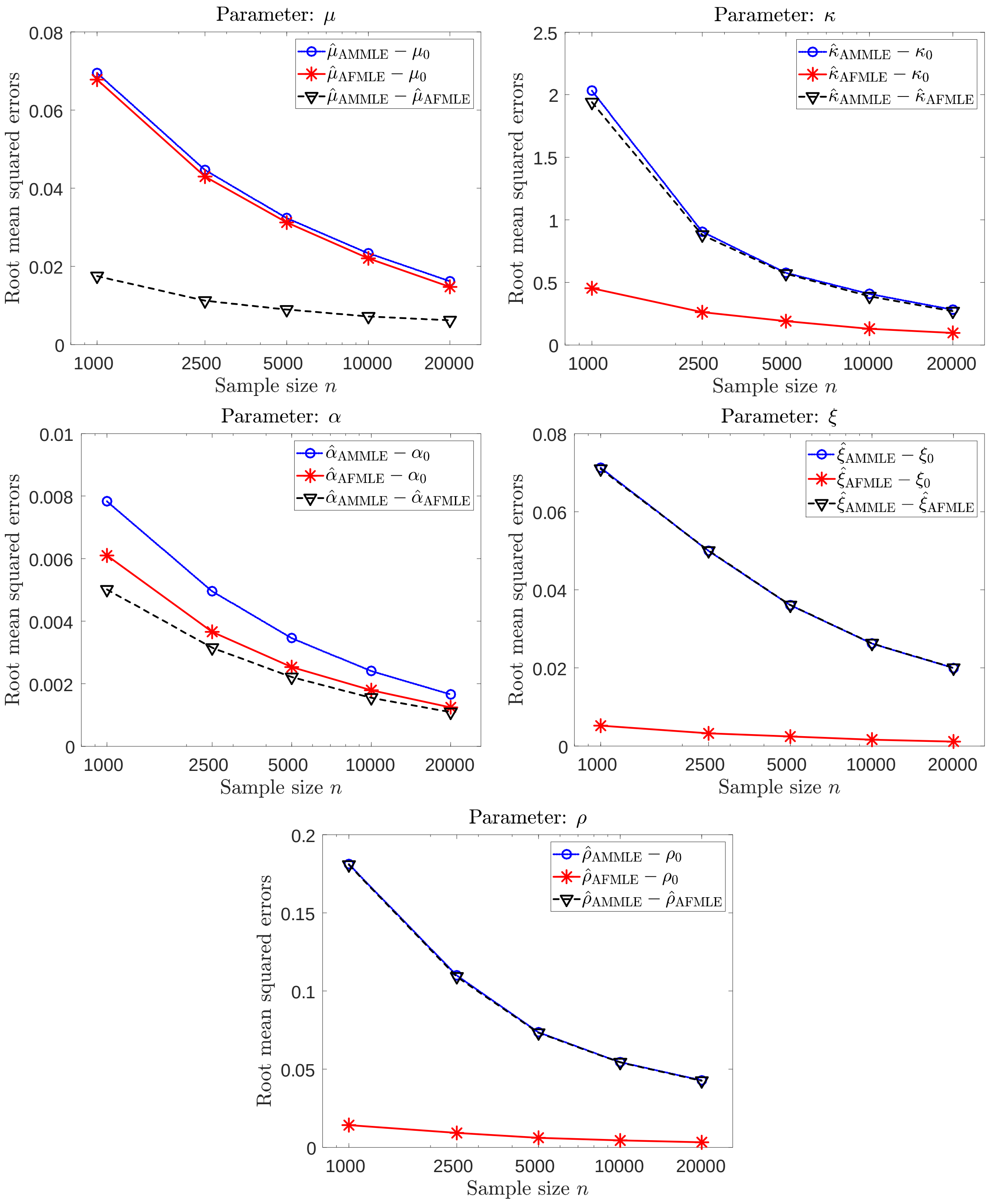

Figure 17: Comparison of AMMLE and AFMLE at weekly frequency

Note: Except for changing the observation frequency to weekly, all the other setting and methods of calculation for producing these five panels are the same as those for producing Figure 16.

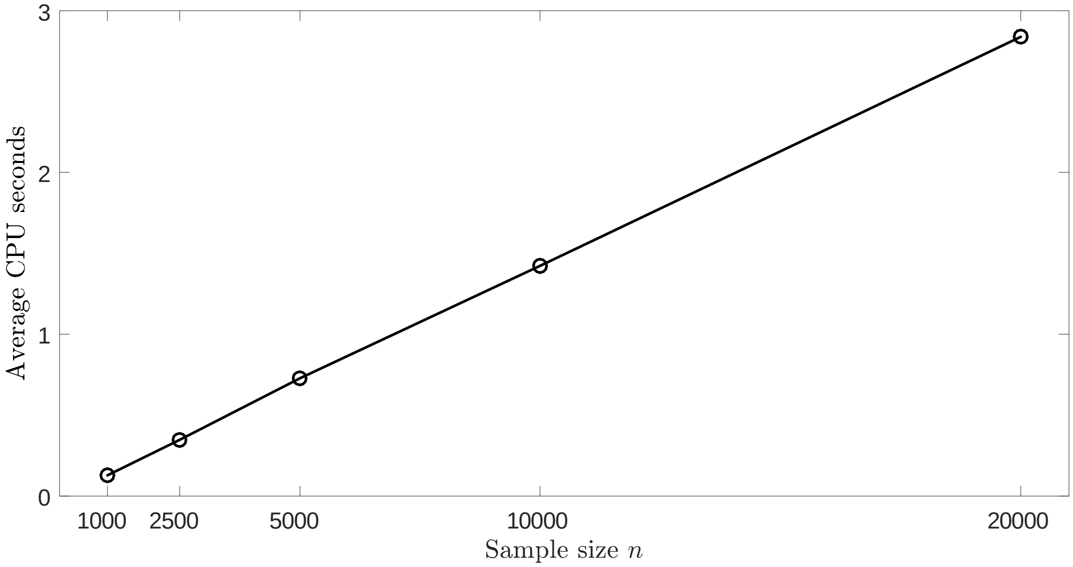

Figure 18: Computational efficiency: Likelihood evaluation

Note: For any sample size n, the scatter point averages the time costs of one-time likelihood evaluation involved in the simulation experiments, with either n daily observations or n weekly observations.

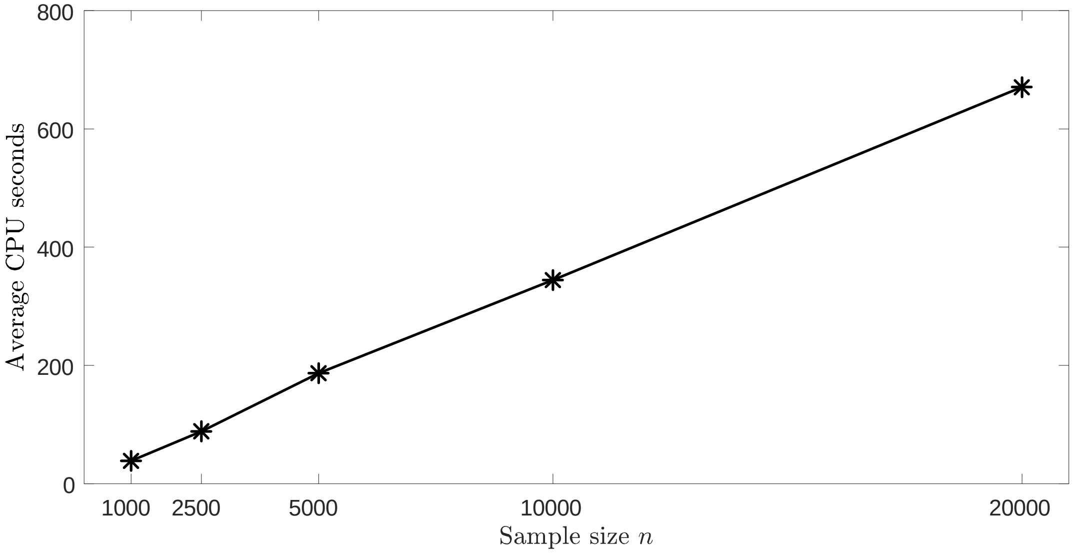

Figure 19: Computational efficiency: Marginal maximum likelihood estimation

Note: For any sample size n, the scatter point averages the time costs of $1,000$ marginal maximum likelihood estimations involved in the simulation experiments – 500 at daily frequency and 500 at weekly frequency.

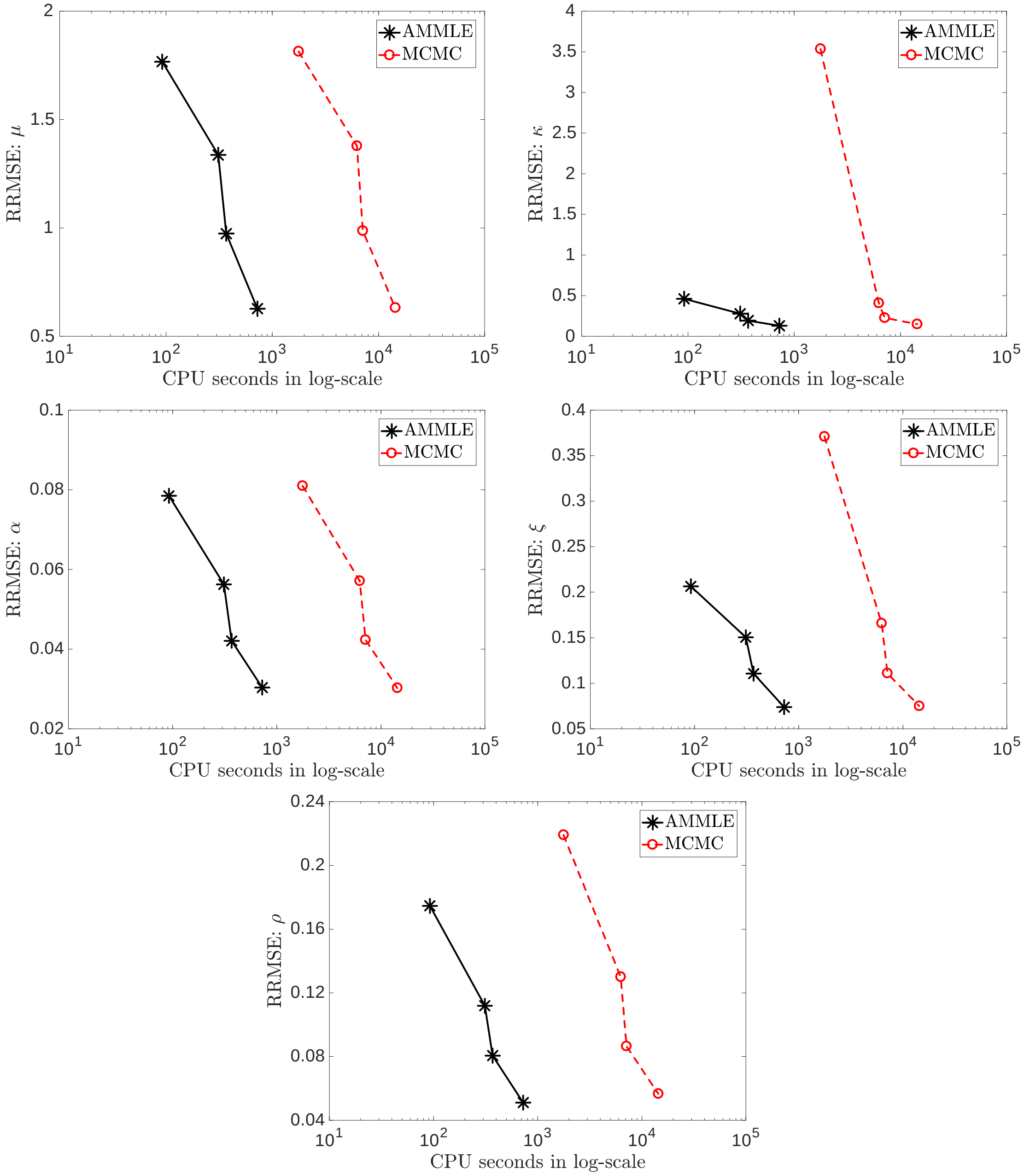

Figure 20: Accuracy vs. computational time: Comparison of methods

Note: This figure compares in Monte Carlo simulations the accuracy (measured by RRMSE in (44)) and computation cost of our approximate marginal MLE (AMMLE) method for estimating the Heston model (39a)–(39b) with that of the conventional Markov Chain Monte Carlo (MCMC) method, by showing the estimation accuracy versus the corresponding time cost. The five panels report the results for the five parameters of the Heston model, whose values are set as $(\mu, \kappa, \alpha, \xi, \rho) = (0.05, 3, 0.1, 0.25, -0.7)$. To report the RRMSE versus the corresponding average CPU time, we repeat 150 simulations for each of the two methods. In each panel, the four starred (resp. circled) points correspond to our AMMLE method (resp. the MCMC method) based on samples with size $n = 2,500$, $5,000$, $10,000$, and $20,000$, respectively.