Electronic Journal of Statistics
Vol. 18 (2024) 3497–3534
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2280

High-frequency volatility estimation and forecasting with a novel Bayesian LGI model

Weiqing Gao[1,2], **Ben Wu**[*1,2] **and Bo Zhang**[*1,2]

[1] *Center for Applied Statistics, Renmin University of China, Beijing, 100872, CN,*
e-mail: gaowq143@ruc.edu.cn*;* wuben@ruc.edu.cn*;* mabzhang@ruc.edu.cn

[2] *School of Statistics, Renmin University of China, Beijing, 100872, CN,*
e-mail: gaowq143@ruc.edu.cn*;* wuben@ruc.edu.cn*;* mabzhang@ruc.edu.cn

Abstract: Volatility modeling is a challenging topic in high-frequency financial data analysis. In this paper, we propose a novel Bayesian framework for modeling and forecasting spot volatility by assuming a latent GARCH structure is embedded into the volatility process at a series of unobserved "anchor" time points, which can well describe the evolving volatility of financial assets in high frequency. We introduce an ideal approximation of latent anchors, which shares similar posterior distribution with true latent anchors. Furthermore, we develop an efficient two-stage inference framework with its corresponding two-stage MCMC sampling algorithm. The simulation study and real data analysis both show our method outperforms the existing alternatives in explanation of latent anchors and the estimation and forecasting of volatility.

MSC2020 subject classifications: Primary 65C05, 60G10; secondary 62F15.

Keywords and phrases: GARCH, high-frequency data, volatility estimation and forecasting, Bayesian inference.

Received August 2023.

1. Introduction

The analysis of volatility is a crucial area of study in financial econometrics and statistics, with numerous practical applications such as asset allocation, option pricing, and risk management. Over the past decades, the increase in accessibility to high-frequency data and the development of relevant modeling tools have led to a better understanding of financial volatility. However, the complexity of volatility dynamics has presented several challenges in modeling high-frequency volatility. Important characteristics of high-frequency financial time series, such as volatility clustering, volatility stationarity, and long memory of volatility, have received significant attention but have not been fully resolved. The dynamics of volatility differ across markets and assets and are influenced by various complex factors, such as public news, dealers' liquidity demand, and

[*] Corresponding author

private information, necessitating a flexible and adaptable modeling framework [5, 24].

Previous research has produced multiple volatility models tailored for high-frequency data. While the classical generalized autoregressive conditional heteroskedastic (GARCH, [18, 10]) model is a robust method for investigating volatility structures in low-frequency financial data, its direct application to high-frequency data is hindered by the problem of parameter inconsistency [4]. Much effort has been devoted to extending the flexibility of GARCH for high-frequency analysis. One such approach is the multiplicative component GARCH model (MC-GARCH, [4, 5, 26, 20]), which applies a GARCH structure to the normalized high-frequency returns after removing daily, diurnal, and seasonal volatility patterns. Another model, the autoregressive conditional duration GARCH (ACD-GARCH, [19, 15]) combines the modeling of observed durations with the GARCH volatility to handle irregularly spaced observations such as tick-by-tick data. Furthermore, the continuous GARCH model [16] and integrated continuous time GARCH model (COGARCH, [39, 46]) extend the conventional discrete-time GARCH model to a continuous-time framework. Nonetheless, the intricate nature of intraday volatility might prove too complex to be accurately estimated with a specific parametric model. [7, 11]. An alternative way is to combine discrete-time GARCH models and realized measures for better utilization of the information inherent in high-frequency data, and thus a better forecasting of the volatility [30, 11, 25]. Empirical studies have shown the good performance of such an idea. [38] took a step forward and proposed the unified GARCH-Itô model (UGI), where a discrete-time GARCH structure is embedded into a continuous-time Itô process for high-frequency financial data at integer time points. The UGI model and its extensions [53, 37] are among the first attempts to provide unified frameworks for both low-frequency and high-frequency volatility modeling and enjoy large flexibility.

In this paper, we propose a Bayesian latent GARCH-Itô model (LGI) for high-frequency data. We assume that the log price of an asset obeys an Itô process, with its volatility process driven by a discrete-time GARCH model defined at a series of latent "anchor" time points. The volatility across two successive anchors exhibits a classical GARCH structure. The anchors are not pre-specified as equally-spaced fixed calendar time points, but are instead latent and should be estimated from the data. By embedding a GARCH structure on such a series of latent anchors, the proposed model enjoys great flexibility in modeling and predicting volatility for high-frequency data. Our work was motivated by the UGI model but has fundamental differences. The UGI model intends to provide a unified modeling framework for both high and low-frequency data. Therefore, its discrete-time GARCH structure is embedded in the volatility process via a series of fixed and known time points, which might be determined ad-hoc in practice. In comparison, we aim to develop a flexible modeling tool for high-frequency data with complex volatility dynamics. Thus we introduce the latent anchor time points to learn the volatility dynamics from the data.

The contributions of this paper can be summarized as follows. First, we propose a sufficiently flexible volatility model for high-frequency data analysis, facilitating the understanding of the complex volatility dynamics of asset returns with high-frequency observations. Second, we propose a framework that offers interpretability for volatility heterogeneity and innovation accumulation in high-frequency data, utilizing a series of unobserved “anchor” time points. This approach enables the forecasting of both spot and integral volatility, providing a comprehensive analysis of the underlying data patterns. Third, to compute more efficiently, we introduce an ideal approximation for latent anchors in the sense that they share similar posterior distributions. Last, we propose a two-stage Bayesian inference procedure and develop the corresponding efficient Markov Chain Monte Carlo sampling algorithm (MCMC) with a birth-death scheme [29, 49] for the posterior computation of the latent anchors and other parameters of interest.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the Bayesian LGI model with the specific prior distribution. Section 3 outlines a methodology for approximating latent anchors and estimates the corresponding convergence rate to the true parameters. In Section 4, we present a two-stage Bayesian inference framework, along with a corresponding two-stage Markov Chain Monte Carlo (MCMC) algorithm. To demonstrate the superiority of our model, we conduct simulation studies in Section 5 and analyze real data obtained from the Shanghai Stock Exchange (SSE) Market in Section 6. Finally, in Section 7, we provide concluding remarks.

2. Method

To predict the asset volatility for risk management, we introduce the latent GARCH-Itô model (LGI) and propose a Bayesian inference procedure on parameters of interest in this section.

Let X_t be the log-price process of an asset defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$. In high-frequency finance, X_t is usually assumed to follow a continuous diffusion process,

$$dX_t = \mu_t dt + \sigma_t dB_t, \tag{1}$$

where μ_t is a drift, B_t is a standard Brownian motion with respect to filtration $\mathcal{F}_t$ and σ_t is the volatility process adapted to $\mathcal{F}_t$. We denote $\tilde{B}_t = B_t - \int_0^t -\mu_s/\sigma_s ds$, Consequently, $\tilde{B}_t$ becomes a Brownian motion with respect to an equivalent probability measure Q, defined as follows according to the Girsanov theorem [48]:

$$dQ = \exp\left\{\int_0^T -\frac{\mu_s}{\sigma_s}dB_s - \frac{1}{2}\int_0^T (\frac{\mu_s}{\sigma_s})^2 ds\right\} dP.$$

And we can deduce that $dX_t = \sigma_t d\tilde{B}_t$ under filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{Q})$, which means that the drift term μ_t will not change the distribution law of X_t fundamentally. Hence, in order to simplify the expression, it is common practice to set the drift term μ_t equal to zero. This simplification has been commonly adopted in prior research studies [1, 47, 38, 53]. To describe the dynamics of σ_t, we assume there exists a series of latent stopping times $\tau_0 < \tau_1 < \ldots$ that cannot be directly observed, and rewrite equation (1) as

$$X_t = X_{\tau_{i-1}} + \int_{\tau_{i-1}}^{t} \sigma_s dB_s, \quad t \in (\tau_{i-1}, \tau_i].$$

In this paper, we name τ_i, $i = 0, 1, \ldots$ "latent anchors", which help us moor a discrete GARCH structure onto the volatility process. Specifically, we assume a GARCH structure is embedded into the volatility process via the latent anchors τ_i, $i = 0, 1, \ldots$ and propose the following LGI model,

$$\begin{cases} X_t = X_{\tau_{i-1}} + Z_t(\tau_{i-1}), \\ \sigma_t^2 = (t - \tau_{i-1})w + \exp\{\gamma(t - \tau_{i-1})\}\sigma_{\tau_{i-1}}^2 + \beta Z_t^2(\tau_{i-1}), \end{cases} \tag{2}$$

for $t \in (\tau_{i-1}, \tau_i]$, $i = 1, \ldots$, where $w \geq 0$, $\gamma < 0$, $\beta > 0$, and $Z_t(\tau_{i-1}) := \int_{\tau_{i-1}}^t \sigma_s dB_s$. In the proposed LGI model (2), the current spot volatility is contributed by three components: the drift term $(t - \tau_{i-1})w$; the spot volatility at the previous anchor $\sigma_{\tau_{i-1}}^2$, weighted by a exponentially decreasing coefficient $\exp\{\gamma(t - \tau_{i-1})\}$; and the innovation accumulated from τ_{i-1} to t. Let p denote the number of latent anchors during $[0, T]$, $\tau_0 = 0$, and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^\top$. We assume τ_i, $i = 1, \ldots$ are arrival times of a Poisson process with intensity $1/\alpha$, and assign the prior for $(\boldsymbol{\tau}, p)$ accordingly as

$$\pi(p, \boldsymbol{\tau}|\alpha) = \alpha^{-p} \exp\left(-\frac{T}{\alpha}\right) I_{\{0 < \tau_1 < \ldots < \tau_p \leq T\}},$$

where $I_{\mathcal{A}}$ is an indicator function of event $\mathcal{A}$.

To show the close connection between the proposed LGI model (2) and the classical discrete-time GARCH model, we give the following assumption on the model parameters.

Assumption 2.1. $w \geq 0, \beta > 0, \gamma < 0, \alpha > 0, (1 - \alpha\beta)(1 - \alpha\gamma) > 1.$

Assumption 2.1 is needed to guarantee the stationary of integral volatility between two consecutive anchors $\int_{\tau_{i-1}}^{\tau_i} \sigma_s^2 ds$, which we will illustrate in the following Proposition.

Proposition 2.1. *Under model (2) and Assumption 2.1, the log-returns between two consecutive anchors $X_{\tau_i} - X_{\tau_{i-1}}$, $i = 1, 2, \ldots$ follow a discrete-time GARCH model,*

$$\begin{cases} X_{\tau_i} - X_{\tau_{i-1}} = Z_{\tau_i}(\tau_{i-1}), \\ E_{\boldsymbol{\xi}}\{Z_{\tau_i}^2(\tau_{i-1})|\mathcal{F}_{\tau_{i-1}}\} = w^g + \gamma^g E_{\boldsymbol{\xi}}\{Z_{\tau_{i-1}}^2(\tau_{i-2})|\mathcal{F}_{\tau_{i-2}}\} + \beta^g Z_{\tau_{i-1}}^2(\tau_{i-2}), \end{cases}$$

where $\boldsymbol{\xi} := (w, \gamma, \beta, \alpha)$, $w^g = w\alpha^2/(1 - \alpha\beta)$, $\gamma^g = 1/(1 - \alpha\gamma)$, $\beta^g = \alpha\beta/(1 - \alpha\beta)(1 - \alpha\gamma)$.

Proposition 2.1 implies that the innovation accumulated over $(\tau_{i-1}, \tau_i]$, $i = 0, 1, \ldots$ can be considered as the discrete-time GARCH innovation. Then, the width of interval $(\tau_{i-1}, \tau_i]$, or the arrival rate of the anchors τ_i, sheds light on the speed of GARCH innovation accumulation. A relatively narrow $(\tau_{i-1}, \tau_i]$, or high arrival rate of τ_i, implies high accumulation speed, and thus fast obtainment of new information from the market. In practice, the innovation accumulation speed usually varies over trading days or over different times within one trading day. The proposed LGI model has the flexibility to capture such variation with its latent anchors.

The proof of Proposition 2.1 is shown in Appendix A.1. With the nonnegativity and stationarity constraints on the GARCH structure, we require the parameters satisfy $w^g \geq 0, 0 < \gamma^g, \beta^g < 1$ and $\gamma^g + \beta^g < 1$, thus $w \geq 0, \beta > 0, \gamma < 0, \alpha > 0$, and $(1-\alpha\beta)(1-\alpha\gamma) > 1$. We assign noninformative flat priors for $\boldsymbol{\xi}$ and σ_0^2, i.e., $\pi(\boldsymbol{\xi}) \propto I_{\{w \geq 0, \beta > 0, \gamma < 0, \alpha > 0, (1-\alpha\beta)(1-\alpha\gamma) > 1\}}$, and $\pi(\sigma_0^2) \propto I_{\{\sigma_0^2 > 0\}}$, which are uniformly distributed on their supports.

3. Posterior distribution

To conduct posterior computation for the proposed LGI model, we rely on parametric modeling of the microstructure noise and a MCMC algorithm. Suppose we observe n log-prices of an asset on a series of observation time points during the observation interval $[0, T]$, denoted as $Y_{t_1}, \ldots, Y_{t_n}$ with $0 = t_0 < t_1 < \ldots < t_n < T$.

3.1. Parametric microstructure noise

Microstructure noise is not negligible in high-frequency data. Following [43], we assume the observed log-prices are contaminated with additive microstructure noise, and the noise is a parametric function of trading information. Specifically, at time t_j, the observed log-price Y_{t_j} is assumed to have the following form,

$$Y_{t_j} = X_{t_j} + h(Z_{t_j}; \boldsymbol{\delta}),$$

where X_{t_j} is the underlying efficient log-price, Z_{t_j} represents observable trading information, such as trading volume, trading type, quoted depth, etc. [27, 50, 3, 34, 13], and $h(.; \boldsymbol{\delta})$ is a parametric function with unknown parameter $\boldsymbol{\delta}$. A typical form of $h(.; \boldsymbol{\delta})$ is linear, where we assume the additive noise is a linear function of observable trading information [44]. Some non-linear form are also proposed in the literature for some particular purpose, such as the segmented model allows asymmetric impacts of buy and sells, and the log model allows concave rates for buys and convex rates for sells [35, 44]. Our framework allows for a very general form of $h(.; \boldsymbol{\delta})$ since we rely on the Metropolis–Hastings algorithm for the sampling of $\boldsymbol{\delta}$.

In simulation and empirical study, we choose a typical function form of noise term,

$$h_{\boldsymbol{\delta}}(\mathcal{R}_{t_j}) = U_{t_j}(\delta_1 + \delta_2 V_{t_j} / \Delta t_j), \tag{3}$$

where V_{t_j} is the trading volume, U_{t_j} is a binary trading type indicator with $U_{t_j} = 1$ representing a buyer-initiated trade and $U_{t_j} = -1$ representing a seller-initiated trade. The rationality of applying equation (3) is based on the fact that one of the most important sources of microstructure noise is the bid-ask spread, which is caused by the trading behaviors of the market participants. A bid (ask) order with a certain volume placed in the market is usually a positive (negative) sign for the future expectation of the asset price, and then the market participants may adjust their strategies accordingly, leading to upward (downward) price movements [28, 40]. Also, this typical function form is usually chosen in previous research [3, 44]. We simply assign a non-informative uniform prior, i.e., $\pi(\boldsymbol{\delta}) \propto 1$ for parameters δ_1 and δ_2.

3.2. Discrete approximation of latent anchors

Let $\mathbf{Y} = (Y_{t_1}, \cdots, Y_{t_n})^\top$ be the observed prices, and $\mathbf{Z} = (Z_{t_1}, \cdots, Z_{t_n})^\top$ be the observed trading information used in the microstructure noise model. Denote all the observed data as $\mathbf{D} = (\mathbf{Y}, \mathbf{Z})$. Then, the joint posterior distribution of the model parameters $\boldsymbol{\theta} = (p, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2)$ given $\mathbf{D}$ is

$$\pi(\boldsymbol{\theta}|\mathbf{D}) \propto \pi(\mathbf{D}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}),$$

where $\pi(\mathbf{D}|\boldsymbol{\theta})$ is an approximated likelihood function defined as a product of a series of normal density, i.e.,

$$\begin{aligned}\pi(\mathbf{D}|\boldsymbol{\theta}) &= \prod_{j=1}^{n} \pi\left(\Delta Y_{t_j}, Z_{t_j} | \boldsymbol{\theta}, \mathcal{F}_{t_{j-1}}\right) \\ &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\left\{ -\frac{\left(\Delta X_{t_j}\right)^2}{2\sigma_{t_{j-1}}^2 \Delta t_j} \right\},\end{aligned}$$

with $\Delta X_{t_j} := X_{t_j} - X_{t_{j-1}}$. However, the likelihood is not directly tractable since for $t_j \in (\tau_{i-1}, \tau_i]$, $\sigma_{t_j}^2$ relies on $Y_{\tau_{i-1}}$, and $Y_{\tau_{i-1}}$ is observable only if the latent anchor τ_{i-1} happens to be an observation time, i.e., $\exists j'$, $\tau_{i-1} = t_{j'}$. In the case that $\tau_{i-1} \neq t_{j'}$, $\forall j'$, we need to integrate out $Y_{\tau_{i-1}}$, which is highly inefficient.

Thus, we propose a discrete approximation scheme for latent anchors. Specifically, we let $\tilde{\tau}_0 = \tau_0 = 0$, and for $i \geq 1$,

$$\tilde{\tau}_i = t_j \quad \text{if and only if} \quad \tau_i \in (t_{j-1}, t_j].$$

$\tilde{\tau}_i$ is then a discrete version of latent anchor τ_i, which only takes value at the observation times $\{t_j\}_{j=1}^n$. To ensure a good approximation of $\boldsymbol{\tau}$, we need the following assumption.

Assumption 3.1. $\inf_i \Delta\tau_i > \sup_j \Delta t_j$ *with* $\Delta t_j = O(1/n)$ *for every* j.

Assumption 3.1 implies the observations are frequent enough that there exists at least one observation between any two consecutive latent anchors, which ensures that $\forall i = 1, \cdots, p-1$,

$$\tilde{\tau}_i \leq \tau_i + \Delta t_j < \tau_i + \Delta\tau_{i+1} = \tau_{i+1} \leq \tilde{\tau}_{i+1},$$

where $\tau_i \in (t_{j-1}, t_j]$. Therefore, the number and the order of the latent anchors remain unchanged after approximation scheme. Also, the interval between any two consecutive observation time points is assumed to be narrow as n is large, which ensures the true τ_i and the approximated $\tilde{\tau}_i$ are close enough. Denote $\mathcal{G} = (t_{j_1-1}, t_{j_1}] \times \cdots \times (t_{j_p-1}, t_{j_p}]$ with $t_{j_1} < \ldots < t_{j_p}$, we can obtain the prior distribution of the approximated latent anchors

$$\begin{aligned}\pi(\tilde{\boldsymbol{\tau}} = \mathbf{t}'|p, \alpha) &= P(\boldsymbol{\tau} \in \mathcal{G}|p, \alpha)\\ &= \int_{\boldsymbol{\tau}\in\mathcal{G}} \alpha^{-p} \exp(-\frac{T}{\alpha}) I_{\{0<\tau_1<\cdots<\tau_p\leq T\}} d\boldsymbol{\tau}\\ &= \alpha^{-p} \exp(-\frac{T}{\alpha}) I_{\{0<t_{j_1}<\cdots<t_{j_p}\leq T\}} \prod_{i=1}^{p} \Delta t_{j_i},\end{aligned}$$

where $\mathbf{t}' = (t_{j_1}, \cdots, t_{j_p})$, $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, \cdots, \tilde{\tau}_p)$. To simplify, we abbreviated $\pi(\tilde{\boldsymbol{\tau}} = \mathbf{t}'|p, \alpha)$ as $\pi(\tilde{\boldsymbol{\tau}}|p, \alpha)$ and denote the prior distribution $\pi(\tilde{\boldsymbol{\theta}}) = \pi(\tilde{\boldsymbol{\tau}}|p, \alpha)\pi(p, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2)$, where $\tilde{\boldsymbol{\theta}} = (p, \tilde{\boldsymbol{\tau}}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2)$. At a high observation frequency, as in a common high-frequency financial application, such approximation is desirable in the sense the bias between the true and approximated posterior distributions can be ignored. We show with the following Theorem 1 the asymptotic property of the bias. We denote the spot volatility corresponding to the approximation as $\tilde{\sigma}_t^2$, with

$$\tilde{\sigma}_t^2 = (t - \tilde{\tau}_{i-1})w + \exp\{\gamma(t - \tilde{\tau}_{i-1})\}\sigma^2_{\tilde{\tau}_{i-1}} + \beta(X_t - X_{\tilde{\tau}_{i-1}})^2.$$

Then, the approximated posterior distribution can be wrote as,

$$\pi(\tilde{\boldsymbol{\theta}}|\mathbf{D}) \propto \pi(\mathbf{D}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}}), \tag{4}$$

To show the convergence of $\tilde{\pi}(\tilde{\boldsymbol{\theta}}|\mathbf{D})$ to $\pi(\boldsymbol{\theta}|\mathbf{D})$, we need the following assumption.

Assumption 3.2. *These exist constants $0 \leq \nu < 1$ and $C > 0$ such that $0 < p \leq Cn^\nu$.*

Assumption 3.2 requires the number of latent anchors p is much smaller than the number of observations n, which is reasonable in high-frequency finance. Our model assumption is more general than UGI model and its extensions, where the number of anchors is assumed to be a constant.

Theorem 3.1. *Under Assumptions 2.1 and 3.1-3.2, we have,*

$$E_{\boldsymbol{\theta}}\left\{\frac{\pi(\tilde{\boldsymbol{\theta}}|\mathbf{D}) - \int_{\boldsymbol{\tau}\in\mathcal{G}} \pi(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\tau}}{\int_{\boldsymbol{\tau}\in\mathcal{G}} \pi(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\tau}}\right\} = o(n^{\nu-1}),$$

where $E_{\boldsymbol{\theta}}$ is the expectation with respect to the sampling distribution of observation data $\mathbf{D}$ under the parameters $\boldsymbol{\theta}$.

Theorem 3.1 ensures the relative bias between the approximated and the true posterior distributions goes to zero as $n \to \infty$. Thus, $\tilde{\boldsymbol{\tau}}$ is an ideal approximation of $\boldsymbol{\tau}$ in the sense that they share similar posterior distributions and the computation on $\tilde{\boldsymbol{\tau}}$ is much more efficient. The proof of Theorem 3.1 is shown in Appendix A.2. In the following section, we will focus on the posterior computation and inferences for the approximated posterior distribution $\pi(\tilde{\boldsymbol{\theta}}|\mathbf{D})$.

4. Posterior computation

In this section, we focus on the posterior inference and sampling of the latent anchors, as well as the spot and integrated volatility. We adopt a two-stage strategy for the inference, and propose the corresponding MCMC algorithm. At the first stage, we obtain the posterior distribution of p and its corresponding point estimate $\tilde{p}$. At the second stage, we sample the other parameters given $p = \tilde{p}$.

4.1. Bayesian inference

The Bayesian inference of model parameters are conducted via a two-stage procedure. At the first stage, we obtain a point estimate of p, such as the maximum a posteriori probability (MAP) estimate, $\tilde{p} = \max_p \pi(p|\mathbf{D})$. At the second stage, we plug in the point estimate of p, and obtain the conditional posterior distribution of the other parameters $\pi(\tilde{\boldsymbol{\tau}}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2|\mathbf{D}, \tilde{p})$, based on which we can make inference on those parameters. In practice, the two-stage procedure facilitates inferences on each of the latent anchors since the number p is fixed at the second stage. We can easily obtain the marginal credible interval given $\tilde{p}$ for each τ_i from the conditional posterior distribution $\pi(\tau_i|\mathbf{D}, \tilde{p})$. The inferences for parameters $\boldsymbol{\xi}$, $\boldsymbol{\delta}$ and σ_0^2 are similar.

Then, we discuss the posterior inferences of the spot and integrated volatility, i.e., $\tilde{\sigma}_t^2$ and $\int_a^b \tilde{\sigma}_t^2 dt$, respectively. Given a time point t_m, we estimate the approximated spot volatility $\tilde{\sigma}_t^2$ with its posterior mean,

$$
\begin{aligned}
E(\tilde{\sigma}_{t_m}^2|\mathbf{D}, \tilde{p}) = {} & E\left\{(t_m - \tilde{\tau}_{p_{t_m}})w|\mathbf{D}, \tilde{p}\right\} \\
& + E\left\{\exp\{\gamma(t_m - \tilde{\tau}_{p_{t_m}})\}\tilde{\sigma}^2_{\tilde{\tau}_{p_{t_m}}}|\mathbf{D}, \tilde{p}\right\} \\
& + E\left\{\beta(X_{t_m} - X_{\tilde{\tau}_{p_{t_m}}})^2|\mathbf{D}, \tilde{p}\right\},
\end{aligned}
\tag{5}
$$

where $\tilde{\tau}_{p_{t_m}}$ is the approximation of $\tau_{p_{t_m}}$, the last latent anchor before t_m. $\tilde{\sigma}^2_{\tilde{\tau}_{p_{t_m}}}$ is obtained recursively.

We estimate the integral volatility over a specific time interval $(a, b]$ with

$$
\tilde{\sigma}^2_{(a,b]} := \sum_{t_j \in (a,b]} \tilde{\sigma}^2_{t_j} \Delta t_j,
$$

Of note, we can also achieve forecasting based on equation (5). Suppose we obtain the log-prices at n discrete observation times t_j, $j = 1, \ldots, n$. To forecast the volatility on discrete time t_j, where $j = n+1, \cdots, m$, we first obtain a series of latent anchors with the discrete probability distribution $\pi(p', \tilde{\boldsymbol{\tau}} = \mathbf{t}_k) = \alpha^{-p'} \exp\{-(t_m - \hat{\tau}_{\tilde{p}})/\hat{\alpha}\} I_{\{\hat{\tau}_{\tilde{p}} < t_{k_1} < \cdots < t_{k_{p'}}\}} \prod_{i=1}^{p'} \Delta t_{k_i}$, where $\mathbf{t}_k = (t_{k_1}, \cdots, t_{k_{p'}})$, $\hat{\alpha}, \hat{\tau}_{\tilde{p}}$ are MAP estimator of α and latent anchor $\tau_{\tilde{p}}$. Then, the last term of equation (5) becomes

$$\begin{aligned}
&E\left\{\beta(X_{t_m} - X_{\tilde{\tau}_{p_{t_m}}})^2|\mathbf{D},\tilde{p}\right\}\\
&= E\left(\beta\int_{t_n}^{t_m}\sigma_s^2 ds|\mathbf{D},\tilde{p}\right) + E\left\{\beta(X_{t_n} - X_{\tilde{\tau}_{p_{t_m}}})^2|\mathbf{D},\tilde{p}\right\}\\
&\approx E\left(\beta\sum_{j=n+1}^{m}\tilde{\sigma}_{t_j}^2\Delta t_j|\mathbf{D},\tilde{p}\right) + (X_{t_n} - X_{\tilde{\tau}_{p_{t_m}}})^2 E(\beta|\mathbf{D},\tilde{p}).
\end{aligned}$$

This implies the future spot volatility can be predicted recursively, based on which we can also obtain a forecasting of future integral volatility.

4.2. Two-stage MCMC algorithm

We develop a two-stage MCMC algorithm based on the Gibbs sampling and Metropolis-Hastings (MH) scheme to obtain the posterior distribution of the latent anchors $p, \tilde{\boldsymbol{\tau}}$, GARCH parameters $\boldsymbol{\xi}$, noise parameters $\boldsymbol{\delta}$ and the initial volatility σ_0^2.

At the first stage, we sample all the parameters of interest using the Gibbs sampling strategy. We employ a birth-death scheme (see, e.g., [29, 49]) to update $(p, \tilde{\boldsymbol{\tau}}(p))$, which involves a birth step to generate a new anchor and increase p, and a death step to discard an old anchor and decrease p. See Algorithm 1 for details. For updating the remaining parameters, a random walk Metropolis-Hastings (MH) algorithm is used. The acceptance rate of the MH algorithm is turned to around 0.2 to ensure a good balance between exploration of the parameter space and convergence to the target distribution. The details of the MCMC algorithm can be found in Appendix B. At the second stage, we fix the number of latent anchors with the MAP estimate using the first stage samples, and update the other parameters of interest with the Gibbs sampling strategy.

4.3. Extensions to jump diffusion and multivariate processes

The prices of financial assets sometimes exhibit drastic fluctuations in a short period of time, which are difficult to fully characterized by traditional diffusion processes [2, 6, 9, 14]. Some researchers propose to use jump diffusion processes in this situation to improve the estimation and prediction of volatility. A jump diffusion process can be written as

$$dX_t = \mu_t dt + \sigma_t dB_t + dJ_t,$$

where J_t represents a "jump term" that captures huge changes of log-prices in a short time. It is obviously that our LGI framework can be extended to accounting for jumps with using the jump diffusion process as the underlying model of the efficient log-price X_t and assigning the jump term an appropriate prior, e.g., the Poisson prior for the arrival of jumps and a Gaussian prior for jump sizes [44]. Another option of handling jumps is to use a thresholding strategy

Algorithm 1 Updating algorithem for $(p, \tilde{\boldsymbol{\tau}})$ with a birth-death scheme

```
Input: learning rate $\epsilon$, latent anchors $(p, \tilde{\tau})$, parameters $\boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2$ and observations $Y, U, V$
Birth step for $p$:
sample $\tau^* \sim Uniform[0, T]$, let $\tilde{\tau}^* = \sum_{j=1}^n t_j I_{\{\tau^* \in (t_{j-1}, t_j]\}}$
if $\tilde{\tau}^* \neq \tilde{\tau}_i$ for $i = 1, \cdots, p$ then
    if there exists $i = 1, \cdots, p$ such that $\tilde{\tau}^* \in (\tilde{\tau}_{i-1}, \tilde{\tau}_i)$ then
        $\tilde{\boldsymbol{\tau}}^* = (\tilde{\tau}_1, \cdots, \tilde{\tau}_{i-1}, \tilde{\tau}^*, \tilde{\tau}_i, \cdots, \tilde{\tau}_p)$
    else if $\tilde{\tau}^* \in (\tilde{\tau}_p, T]$ then
        $\tilde{\boldsymbol{\tau}}^* = (\tilde{\tau}_1, \cdots, \tilde{\tau}_p, \tilde{\tau}^*)$
    end if
    sample $u \sim Uniform[0, 1]$, let $\rho = \frac{\pi(p+1, \tilde{\boldsymbol{\tau}}^* | \xi, \delta, \sigma_0^2, \mathbf{D})}{\pi(p, \tilde{\boldsymbol{\tau}} | \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2, \mathbf{D})}$
    if $u < min(1, \rho)$ then
        update $p = p + 1$, $\tilde{\boldsymbol{\tau}} = \tilde{\boldsymbol{\tau}}^*$
    end if
end if
Death step for $p$:
sample $i^*$ with $P(i^* = i) = 1/p, i = 1, \cdots, p$
$\tilde{\boldsymbol{\tau}}^* = (\tilde{\tau}_1, \cdots, \tilde{\tau}_{i^*-1}, \tilde{\tau}_{i^*+1}, \cdots, \tilde{\tau}_p)$
sample $u \sim Uniform[0, 1]$, let $\rho = \frac{\pi(p-1, \tilde{\boldsymbol{\tau}}^* | \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2, \mathbf{D})}{\pi(p, \tilde{\boldsymbol{\tau}} | \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2, \mathbf{D})}$
if $u < min(1, \rho)$ then
    update $p = p - 1$, $\tilde{\boldsymbol{\tau}} = \tilde{\boldsymbol{\tau}}^*$
end if
Update $\tilde{\boldsymbol{\tau}}$:
for $i = 1, 2, \cdots, p$ do
    sample $\tau_i^* \sim N(\tilde{\tau}_i, \epsilon^2)$, let $\tilde{\tau}_i^* = \sum_{j=1}^n t_j I_{\{\tilde{\tau}_i^* \in (t_{j-1}, t_j]\}}$
    sample $u \sim Uniform[0, 1]$, let $\rho = \frac{\pi(\tilde{\tau}_i^* | \tilde{\boldsymbol{\tau}}_{(-i)}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2, \mathbf{D})}{\pi(\tilde{\tau}_i | \tilde{\boldsymbol{\tau}}_{(-i)}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2, \mathbf{D})}$
    if $u < min(1, \rho)$ then
        update $\tilde{\tau}_i = \tilde{\tau}_i^*$
    end if
end for
```

to screen out the jumps from the observed prices before any further analysis, which has shown its advantages in many empirical studies [45, 41, 42]. In our empirical study section, we follow [23] to screen out jumps by setting a specific threshold for the log returns, and then build the LGI model on the jump-free log return series. We focus on modeling the volatility in this paper, and leave the joint model of volatility and jumps for future works.

In addition, our model can be extended to handling multivariate asset price series analysis by borrowing ideas from the latent factor model. We can use a similar strategy as the factor GARCH-Itô model proposed by [36], which puts a GARCH-Itô structure on latent factors to model the volatility matrix for multivariate series. Specifically, we assume the multivariate price series are driven by r common latent factors, and the volatility matrix admits an eigendecomposition, with the instantaneous eigenvalues $\lambda_{t,j}$, $j = 1, \ldots, r$ following the LGI structure, i.e.,

$$\lambda_{t,j} = (t - \tau_{i-1}) w_j + \exp\{\gamma_j (t - \tau_{i-1})\} \lambda_{\tau_{i-1}, j} + \sum_{l=1}^{r} \beta_{j,l} \left\{ \int_{\tau_{i-1}}^{t} \sqrt{\lambda_{s,l}} dW_{s,l} \right\}^2, \quad (6)$$

for $j = 1, \dots, r$. Another challenge in extension of LGI model to multivariate cases is the non-synchronicity in multivariate observations. Fortunately, there are already synchronizing methods developed in literature [1, 22, 21, 55]. We can rely on a global refresh time method [8] to synchronize multivariate observations from all the assets, then evaluate the instantaneous eigenvalues on the synchronized times $\tilde{t}$ to obtain a LGI structure for $\lambda_{\tilde{t},j}$, $j = 1, \dots, r$. For posterior inferences, we may use the same synchronized times $\tilde{t}$ to calculate the realized covariance matrix on interval $(\tau_{i-1}, \tau_i]$, denoted as $\mathbf{\Sigma}_i$, and then find an approximation of the integral term in equation (6) with the eigendecomposition of $\mathbf{\Sigma}_i$. Alternatively, we can adopt a pairwise refresh time method for calculating each off-diagonal element of the realized covariance matrix since it only involves observations from two specific assets, which may leads to a better approximation of the integral term.

5. Simulation

In this section, we begin by generating data using a toy model to assess the estimation performance of all parameters, including the latent anchors τ_i, at varying observation frequencies. Next, to evaluate the volatility estimation and forecasting performance of the LGI model with both true model and model with misspecification, we consider three scenarios with different data-generating mechanisms: data generated with 1) LGI model, 2) the Ornstein-Uhlenbeck model [56, 52, 51], and 3) the Heston model [31], where 2) and 3) are well-known models for interpretation of the volatility evolution of financial assets.

5.1. Parameter settings

In Scenario 1, we generate latent prices with LGI model. We set $\Delta\tau_i \sim \exp(1)$ and $w = 0.075, \gamma = -2.5, \beta = 0.35, \sigma_0^2 = 0.02$.

In Scenario 2, we generate latent prices with an Ornstein-Uhlenbeck model,

$$\begin{aligned} dX_t &= \sigma_t dB_t, \\ dlog(\sigma_t) &= \theta_\sigma(v_\sigma - log(\sigma_t))dt + v dB_t^\sigma, \end{aligned}$$

and in Scenario 3, we generate latent prices with a Heston model,

$$\begin{aligned} dX_t &= \sigma_t dB_t, \\ d\sigma_t^2 &= \theta_\sigma(v_\sigma - \sigma_t^2)dt + v\sigma_t dB_t^\sigma. \end{aligned}$$

Both models are constructed with two related stochastic diffusion processes: the price process and the volatility process. B_t, B_t^σ are two correlated standard Brownian motions with correlation $\rho_1 = -0.5$. Under the Heston model, v_σ and v can be explained as the long-term price variance and the volatility of volatility (VOV) process, respectively; and θ_σ is the rate of reversion to the long-term price variance, which determines the relative weights of the current variance and the

Table 1
Parameter settings of Scenario 2 and Scenario 3.

	Scenario 2				Scenario 3			
	SL	SH	FL	FH	SL	SH	FL	FH
θ_σ	0.7	0.7	1.4	1.4	0.7	0.7	1.4	1.4
v	0.1	0.5	0.1	0.5	0.1	0.2	0.1	0.2
v_σ	-1.5	-1.5	-1.5	-1.5	0.1	0.1	0.1	0.1

log-term variance of the prices. The explanation of the parameters are similar under the Ornstein-Uhlenbeck model but on the log scale. We set each 4 settings for Scenario 2 and Scenario 3 as Table 1, where SL, SH, FL, FH represent slow reversion and low VOV, slow reversion and high VOV, fast reversion and low VOV, fast reversion and high VOV respectively.

We consider equally spaced observation time points $0 = t_0 < t_1 < \ldots < t_n = T$ with $T = 100$ denoting 100 consecutive trading days, and $\Delta := t_{j+1} - t_j$ denoting the observation frequency. Three different frequencies are investigated with $\Delta = 1/500$, $\Delta = 1/1500$ and $\Delta = 1/5000$. We generate the observed log-prices $\{Y_{t_j}\}_{j=1}$ with the parametric noise model as in equation (3), with the parameters of noise term set as $\delta_1 = 1.5 \times 10^{-3}$ and $\delta_2 = 3 \times 10^{-10}$. The corresponding signal-to-noise ratio in these settings ranges from 0.60 to 0.99, which is similar with our observations in the empirical study. We simulate 50 datasets for each simulation setting.

For the implementation of the first-stage MCMC algorithm, we draw 5,000 samples with random initializations and discard the first 2,000 samples as burn-in. Then we draw 5,000 samples with the second-stage MCMC algorithm.

To evaluate the convergence of the algorithm, we adopt the Gelman-Rubin diagnostic with 10 repetitions of the experiment. The details of the convergence diagnostic results can be found in Appendix C.1.

5.2. Estimation and forecasting results

We use the samples from the first 99 trading days to evaluate the estimation performance of the methods, and make one-day-ahead forecasting of the integral volatility on the 100-th trading day for the evaluation of forecasting. Table 2 reports the estimation of parameters $w, \gamma, \beta, \alpha, \sigma_0^2, \delta_1, \delta_2$ with the posterior average (Mean), standard deviation (Std), confidence interval (C.I.), effective size (Eff) and accept rate (AC) under different observation frequency $\Delta = 1/500, 1/1500, 1/5000$. The estimation error of the parameters primarily decreases as the number of intraday observations increases, demonstrating the favorable limited sample property of the model. To evaluate the estimation of latent anchors, we denote the estimation error as

$$\frac{1}{p} \sum_{i=1}^{p} \min_j |\hat{\tau}_j - \tau_i|,$$

Table 2
Estimation of LGI model parameters over 50 simulation runs.

	True		Mean	Std	C.I.	Eff	AC
w	0.075	$\Delta = 1/500$	0.074	0.002	[0.069,0.077]	109.825	0.200
		$\Delta = 1/1500$	0.074	0.002	[0.073,0.077]	124.186	0.198
		$\Delta = 1/5000$	0.074	0.002	[0.069,0.075]	89.130	0.195
γ	-2.5	$\Delta = 1/500$	-2.485	0.099	[-2.672,-2.298]	143.290	0.201
		$\Delta = 1/1500$	-2.481	0.073	[-2.602,-2.314]	142.657	0.201
		$\Delta = 1/5000$	-2.439	0.095	[-2.543,-2.229]	93.705	0.201
β	0.35	$\Delta = 1/500$	0.344	0.018	[0.308,0.376]	129.111	0.201
		$\Delta = 1/1500$	0.356	0.027	[0.327,0.428]	115.692	0.199
		$\Delta = 1/5000$	0.368	0.047	[0.338,0.488]	79.939	0.198
α	1	$\Delta = 1/500$	1.120	0.184	[0.796,1.539]	573.162	0.207
		$\Delta = 1/1500$	1.078	0.159	[0.778,1.448]	734.664	0.207
		$\Delta = 1/5000$	0.987	0.150	[0.774,1.270]	690.839	0.207
$\delta_1(\times 10^{-3})$	1.5	$\Delta = 1/500$	1.509	0.122	[1.282,1.737]	391.430	0.201
		$\Delta = 1/1500$	1.504	0.022	[1.463,1.544]	542.633	0.201
		$\Delta = 1/5000$	1.501	0.005	[1.492,1.508]	221.814	0.197
$\delta_2(\times 10^{-10})$	3	$\Delta = 1/500$	2.977	0.200	[2.585,3.375]	407.511	0.201
		$\Delta = 1/1500$	2.995	0.023	[2.947,3.043]	807.778	0.201
		$\Delta = 1/5000$	2.999	0.002	[2.997,3.003]	1015.182	0.197
σ_0^2	0.02	$\Delta = 1/500$	0.021	0.005	[0.014,0.032]	1003.778	0.204
		$\Delta = 1/1500$	0.020	0.003	[0.015,0.026]	1012.271	0.201
		$\Delta = 1/5000$	0.020	0.002	[0.017,0.022]	997.160	0.202

Table 3
The estimation error and the length of 90%, 95% posterior interval for latent anchors, shown as mean (standard deviation) over 50 simulation runs.

	Err	90% interval	95% interval
$\Delta = 1/500$	**0.115(0.021)**	0.884(0.234)	1.018(0.264)
$\Delta = 1/1500$	**0.075(0.019)**	0.424(0.118)	0.490(0.135)
$\Delta = 1/5000$	**0.059(0.036)**	0.240(0.063)	0.277(0.072)

where $\hat{\tau}_i$ is the MAP estimation of latent anchors introduced in Section 4.1. Besides, we also obtain the average length of posterior interval for latent anchors,

$$\frac{1}{\tilde{p}}\sum_{i=1}^{\tilde{p}} |I_{\tau_i}(\alpha)|$$

where $I_{\tau_i}(\alpha)$ is for the α-interval for each latent anchors τ_i under the conditional posterior distribution $\pi(\tau_i|\mathbf{D}, \tilde{p})$, $|\cdot|$ denotes the length of the interval. Table 3 presents both the estimation error of latent anchors under MAP estimator and the average length of 90% and 95% posterior intervals for the latent anchors. From the decreasing error and interval length with the increasing intraday observations, we can further highlight the model's ability to handle limited samples.

To evaluate the estimation and forecasting performance of integral volatility, we implement two existing volatility modeling methods: UGI model and estimated-price realized volatility method (ERV, [44]) for comparison. The es-

Table 4

The estimation and forecasting errors of the integral volatility under LGI, UGI and ERV for both simulation scenarios, shown as mean (standard deviation) over 50 simulation runs.

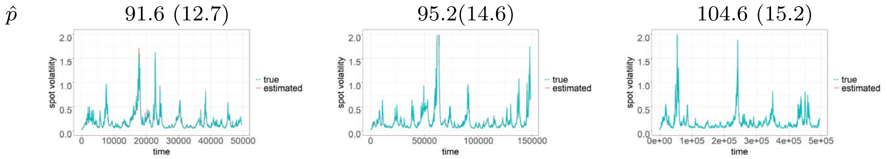
		Estimation			Forecasting		
Scenario 1		$\Delta = 1/500$	$\Delta = 1/1500$	$\Delta = 1/5000$	$\Delta = 1/500$	$\Delta = 1/1500$	$\Delta = 1/5000$
	LGI	**0.043(0.004)**	**0.026(0.005)**	0.019 (0.009)	**0.289(0.257)**	**0.311(0.218)**	**0.185(0.134)**
	UGI	0.534 (0.084)	0.544(0.119)	0.512(0.072)	0.546 (0.581)	0.610 (0.693)	0.383 (0.330)
	ERV	0.052 (0.004)	0.030 (0.002)	**0.016(0.001)**	0.456 (0.519)	0.532 (0.639)	0.367 (0.232)
Scenario 2		$\Delta = 1/500$	$\Delta = 1/1500$	$\Delta = 1/5000$	$\Delta = 1/500$	$\Delta = 1/1500$	$\Delta = 1/5000$
SL	LGI	0.068 (0.011)	0.055 (0.011)	0.048 (0.012)	**0.117(0.079)**	**0.096(0.061)**	**0.109(0.071)**
	UGI	0.106 (0.009)	0.100 (0.009)	0.097 (0.010)	0.118 (0.089)	0.099 (0.071)	0.112 (0.077)
	ERV	0.051 (0.004)	**0.029(0.002)**	0.016 (0.001)	0.122 (0.093)	0.097 (0.071)	0.118 (0.083)
SH	LGI	0.200 (0.028)	0.187 (0.025)	0.166 (0.021)	**0.592(0.648)**	**0.635(0.582)**	**0.506(0.534)**
	UGI	0.745 (0.100)	0.763 (0.103)	0.748 (0.083)	0.795 (0.800)	0.866 (1.035)	0.752 (0.809)
	ERV	**0.089(0.008)**	**0.079(0.007)**	**0.074(0.006)**	0.707 (0.793)	0.675 (0.813)	0.573 (0.552)
FL	LGI	0.058 (0.007)	0.045 (0.008)	0.040 (0.011)	0.091 (0.057)	0.085 (0.064)	**0.071(0.050)**
	UGI	0.078 (0.007)	0.075 (0.008)	0.076 (0.007)	**0.079(0.061)**	**0.079(0.067)**	0.083 (0.062)
	ERV	**0.051(0.004)**	**0.029(0.002)**	**0.016(0.001)**	0.105 (0.068)	0.096 (0.074)	0.089 (0.067)
FH	LGI	0.144 (0.016)	0.132 (0.017)	0.122 (0.014)	**0.401(0.400)**	**0.445(0.349)**	**0.400(0.341)**
	UGI	0.464 (0.059)	0.457 (0.053)	0.466 (0.055)	0.527 (0.601)	0.601 (0.619)	0.561 (0.597)
	ERV	**0.082(0.007)**	**0.068(0.006)**	**0.064(0.005)**	0.577 (0.761)	0.524 (0.518)	0.663 (0.692)
Scenario 3		$\Delta = 1/500$	$\Delta = 1/1500$	$\Delta = 1/5000$	$\Delta = 1/500$	$\Delta = 1/1500$	$\Delta = 1/5000$
SL	LGI	0.089 (0.016)	0.070 (0.013)	0.060 (0.013)	**0.154(0.087)**	**0.184(0.149)**	**0.138(0.110)**
	UGI	0.162 (0.016)	0.161 (0.016)	0.163 (0.013)	0.167 (0.121)	0.193 (0.138)	0.164 (0.134)
	ERV	**0.051(0.004)**	**0.029(0.003)**	**0.016(0.001)**	0.167 (0.141)	0.187 (0.149)	0.163 (0.127)
SH	LGI	0.124 (0.024)	0.116 (0.020)	0.097 (0.013)	**0.338(0.296)**	0.351 (0.349)	**0.316(0.343)**
	UGI	0.362 (0.043)	0.358 (0.049)	0.371 (0.050)	0.399 (0.442)	**0.330(0.345)**	0.484 (0.456)
	ERV	**0.053(0.004)**	**0.030(0.001)**	**0.016(0.001)**	0.414 (0.375)	0.356 (0.333)	0.449 (0.395)
FL	LGI	0.077 (0.012)	0.064 (0.012)	0.054 (0.011)	**0.120(0.084)**	**0.122(0.106)**	**0.131(0.096)**
	UGI	0.119 (0.012)	0.120 (0.011)	0.119 (0.011)	0.139 (0.096)	0.137 (0.110)	0.148 (0.133)
	ERV	**0.051(0.004)**	**0.029(0.002)**	**0.016(0.001)**	0.132 (0.119)	0.152 (0.119)	0.159 (0.133)
FH	LGI	0.150 (0.019)	0.084 (0.015)	0.080 (0.013)	**0.221(0.212)**	**0.188(0.132)**	**0.245(0.168)**
	UGI	0.236 (0.024)	0.241 (0.027)	0.246 (0.027)	0.230 (0.186)	0.214 (0.151)	0.256 (0.214)
	ERV	**0.052(0.004)**	**0.029(0.003)**	**0.016(0.001)**	0.260 (0.169)	0.214 (0.160)	0.269 (0.229)

timation or forecasting error of each method is measured as follows,

$$err = \sum_{d \in \mathcal{D}} \frac{1}{|\mathcal{D}|} err_d = \sum_{d \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \frac{\left|\hat{h}_d - \int_{d-1}^{d} \sigma_s^2 ds\right|}{\int_{d-1}^{d} \sigma_s^2 ds},$$

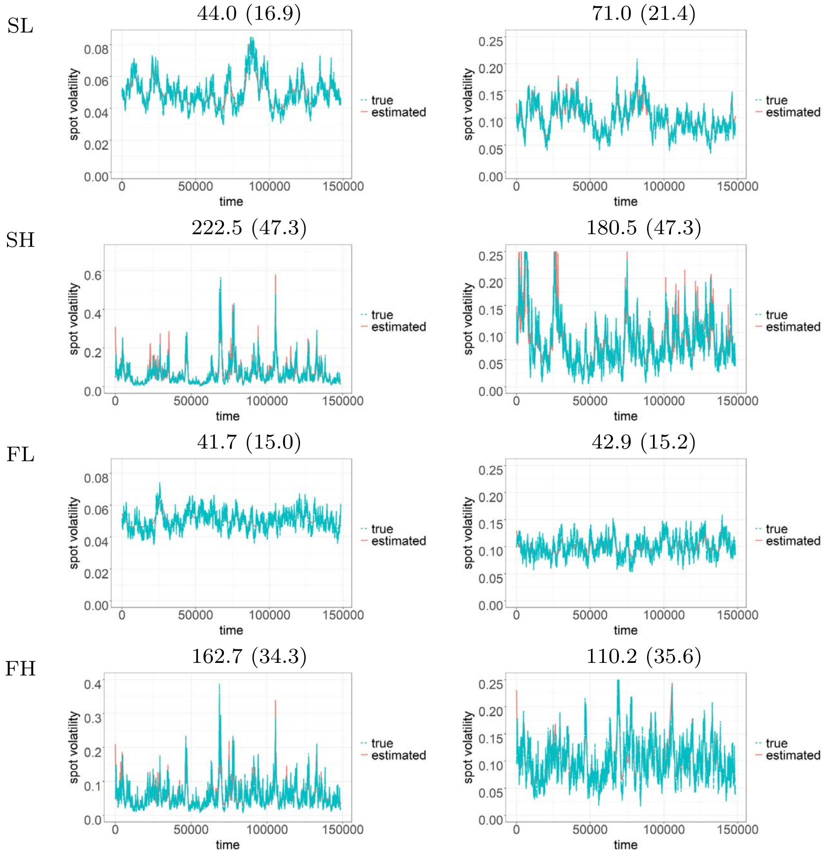
where $\hat{h}_d$ is the estimation or forecasting of the integral volatility on the d-th day, and $\mathcal{D}$ is the index set containing all the trading days evaluated. Both LGI model and UGI model provide the one-day-ahead forecasting of the integral volatility, and we use the $(d-1)$-th estimated realized volatility as the forecasting of the d-th day integral volatility under the ERV method. Table 4 shows the estimation and forecasting errors of the LGI model, UGI model and ERV method over 50 simulation runs under both scenarios. Under the true model in Scenario 1, it is observed that the LGI model exhibits superior predictive accuracy compared to the UGI and ERV models. Additionally, in terms of estimation accuracy, the LGI model displays significantly better performance than the UGI model and shows comparable or even better results than the ERV model, particularly when the observation frequency is low. In scenarios where model misspecification occurs (Scenario 2 and 3), the LGI model demonstrates superior performance in estimation compared to the UGI model, owing to its more flexible structure that aids in fitting under model misspecification. The LGI model also achieves similar estimation results to the ERV model in most settings, which highlights its flexibility. In terms of forecasting, all three methods have comparable performance with LGI being slightly better than UGI and ERV.

Figure 1-2 show the estimated spot volatility and the number of latent anchors under the LGI model for Scenario 1,2,3. The proposed model exhibits a high level


$\hat{p}$ 91.6 (12.7) 95.2(14.6) 104.6 (15.2)

(a) $\Delta = 1/500$ (b)$\Delta = 1/1500$ (c) $\Delta = 1/5000$

FIG 1. *True volatility and volatility estimated by the LGI model with* $\Delta = 1/500, 1/1500, 1/5000$ *under toy model. The mean (standard deviation) of the number of latent anchors over 50 simulation runs are shown on the top of each plot.*


SL 44.0 (16.9) 71.0 (21.4)

SH 222.5 (47.3) 180.5 (47.3)

FL 41.7 (15.0) 42.9 (15.2)

FH 162.7 (34.3) 110.2 (35.6)

(a) Scenario 1 (b) Scenario 2

FIG 2. *True volatility and volatility estimated by the LGI model with* $\Delta = 1/1500$. *Figure(a) and Figure(b) shows the single-run results under Scenario 2 and Scenario 3 respectively. The mean (standard deviation) of the number of latent anchors over 50 simulation runs are shown on the top of each plot.*

of accuracy in capturing major spot volatility patterns under all settings, even in the presence of model misspecification, particularly in scenarios with high VOV (SH and FH) and more latent anchors. This outcome suggests that the proposed model has a flexible structure capable of accommodating different scenarios effectively. When the VOV is low, the mean-reversion term becomes relatively larger, resulting in a more pronounced model misspecification. Nevertheless, the LGI model can still capture the primary trends of the volatility process in such scenarios, albeit with some loss of short-term disturbances. These results remain consistent with different observation frequencies of $\Delta = 1/500$ and $\Delta = 1/5000$, as demonstrated in Appendix C.2.

Besides, the computation flexibility of our methods is $O(T/\Delta)$. Specifically, the average computation time under Scenario 1 for the intraday frequency $\Delta = 1/500, 1/1500, 1/5000$ is 9.948, 29.295, and 96.799, respectively. This outcome suggests that the proposed methods can handle large datasets and provide accurate results in a reasonable amount of time, even with high-frequency data.

6. Empirical study

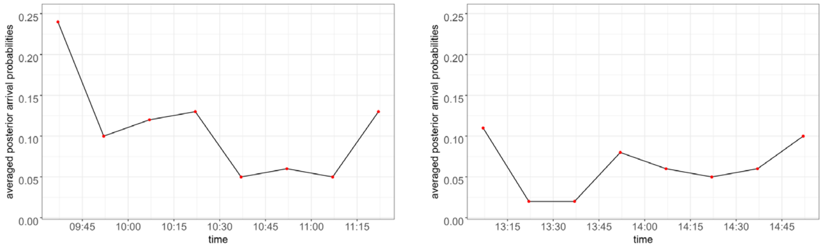
In this section, we apply the proposed method to analyze the asset price volatility using high-frequency financial data collected in the Shanghai Stock Exchange (SSE). The data is obtained from the Wind database.

6.1. Data description

We collect the tick-by-tick observed log prices trading types, and trading volumes with $T = 100$ trading days from 1/12/2014 to 29/4/2015 for three stocks: Shanghai Pudong Development Bank (SPD BANK), Guangzhou Baiyun International Airport Company (CAN) and China Minsheng Banking (CMBC). In the following, we focus on the analysis of SPD BANK, and show the results of CAN and CMBC in Appendix D.2. The observation frequency Δ we study is one second, leading to $14,400$ intraday observations per day. The per-second observed log price at t_j, denoted as Y_{t_j}, is taken as the average of the tick-by-tick observed log prices on $[t_j, t_{j+1})$. The per-second trading volume, denoted as V_{t_j}, is taken as the absolute value of the difference between the total buying and selling volumes on $[t_j, t_{j+1})$. The per-second trading type U_{t_j} is determined by the total buying and selling volumes on $[t_j, t_{j+1})$, and takes $+1$ when the buying volume is greater than the selling volume and -1 otherwise. Following [23], we screen out price jumps, i.e., abnormally large per-second log returns that are unable to be characterized by the continuous Itô diffusion process [7, 9] by setting any log return with its absolute value larger than a specific threshold to be zero.

6.2. Model inference

As introduced in Section 4.1, we mainly focus on the analysis of the latent anchors, the integral and spot volatility. Inferences on the other parameters are



Mid-Morning Market Mid-Afternoon Market

FIG 3. *The averaged posterior arrival probabilities for latent anchors over the time interval $\mathcal{T}_j, j = 1, \cdots, M$ in the whole $T = 100$ trading days for SPD BANK.*

reported in Appendix D.1. For the implementation of the LGI model, we draw 25,000 samples from the first-stage Markov chain and disregard the first 15,000 draws as burn-in samples. Then, we draw 25,000 samples from the second-stage Markov chain with the number of latent anchors fixed.

6.2.1. Inference of latent anchors

We divide one trading day into $M = 16$ 15-minute time intervals $\mathcal{T}_j, j = 1, \cdots, M$, and investigate the posterior probability for the point estimate of latent anchors $Prob_{\mathcal{T}_j}$ on each of them,

$$Prob_{\mathcal{T}_j} = \frac{1}{T} \sum_{i=1}^{\tilde{p}} I_{A_{i,j}},$$

where the set $A_{i,j} = \{\exists i, \hat{\tau}_i \in \mathcal{T}_{i,j}\}$, $\mathcal{T}_{i,j}$ is the $\mathcal{T}_j$ interval in the i-th trading day.

Figure 3 shows the averaged posterior probability for the point estimate of latent anchors over the $T = 100$ trading days for both mid-morning and mid-afternoon markets. The latent anchors have a relatively higher average probability in the first hour ($\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ and $\mathcal{T}_4$) right after the opening and the last 15-minute before the closing of the market ($\mathcal{T}_{16}$), especially in the within the 15-minute periods ($\mathcal{T}_1$). Besides, within the 15-minute both before the closing of mid-morning and after the opening of mid-afternoon markets ($\mathcal{T}_8$ and $\mathcal{T}_9$), the averaged posterior probability for the point estimate of latent anchors exceeds other time. The result from Figure 3 indicates the proposed LGI model finds fast GARCH innovation accumulation during the opening and closing of the markets, which is consistent with empirical findings in the existing literature. In the stock market, the public information released overnight and during lunch break leads to the surge of volatility in the opening period of mid-morning and mid-afternoon markets respectively [33, 32]. Also, [17, 12] pointed out that the trading volume during the opening and closing time is usually higher than in other periods, along with much more disclosure of private information.

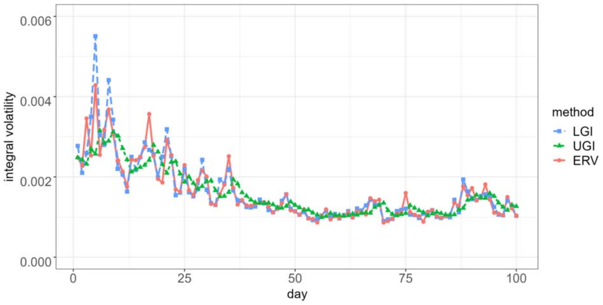


Fig 4. *The estimation of integral volatility for SPD BANK under LGI, UGI, ERV over the 100 trading days from 1/12/2014 to 29/4/2015.*

Table 5
The averaged forecasting error of SPD BANK for integral volatility under LGI, UGI and ERV model.

	LGI	UGI	ERV
RMSE ($\times 10^{-4}$)	2.149	2.470	2.572
MAE ($\times 10^{-4}$)	1.637	1.971	2.012

6.2.2. *Analysis of integral volatility*

Figure 4 presents the integral volatility for the 100 trading days estimated with the LGI model, UGI model, and ERV method. As a semiparametric model, ERV has few assumptions on the volatility process and thus can be regarded as a benchmark of integral volatility estimation. The three methods yield similar results, which implies the stability of the LGI model. To evaluate the forecasting of the integral volatility, we employ a sliding window-based model evaluation procedure. Specifically, we train the models using the high-frequency observations on a sliding window $W_d = [d - L - 1, d - 1)$ of width L trading days and make one-day-ahead forecasting for the integral volatility on the d-th day over $[d-1, d)$. We take $L = 80$, and construct $q = 20$ windows (for $d = 81, \ldots, 100$) over the 100 trading days investigated. Table 5 shows the averaged forecasting error of SPD BANK for integral volatility under LGI, UGI and ERV model with error measurement RMSE and MAE denoted as

$$RMSE = \sqrt{\frac{1}{q}\sum_{d=L+1}^{L+q}(\hat{h}_d - ERV_d)^2},$$

$$MAE = \frac{1}{q}\sum_{d=L+1}^{L+q}|\hat{h}_d - ERV_d|,$$

where $\hat{h}_d$ is the one-day-ahead forecasting of the integral volatility over $[d-1, d)$ using observations on the window W_d. As a result, the LGI model has a better performance on integral volatility forecasting compared with its competitors. In conclusion, the LGI model is a highly effective tool for both integral volatility estimation and forecasting.

6.2.3. Analysis of spot volatility

To evaluate the performance of the proposed LGI model in terms of spot volatility analysis, we consider several alternative high-frequency volatility models including intraday GARCH [20], intraday exponential GARCH (intraday EGARCH, [20]), auto-regressive stochastic volatility (ARSV, [54]) and Heston models [31]. To eliminate the impact of microstructure noise, we assume the noise has the same parametric form as in equation (3) for all the alternative modes, and apply a two-stage strategy to estimate the parameters. We first estimate the noise parameters with the ordinary least squares (OLS) method proposed in [44], and then solve the volatility models using the estimated efficient prices after the removal of noise terms.

We use the log-likelihood and the Bayesian information criterion (BIC) to quantify model fitting. According to the results presented in Table 6, it can be observed that the LGI model achieved the highest log-likelihood and the lowest BIC among the evaluated models for the $T = 100$ trading days. This indicates that the LGI model outperforms the other models in accurately estimating spot volatility. In order to conduct a thorough evaluation of the forecasting performance of the models, we employed the sliding window procedure introduced in the previous section, utilizing the log predictive score and BIC as measures of model performance. The mean and standard deviation of the log predictive score and BIC for the one-day-ahead spot volatility forecasting with the 20 sliding windows using the five models are presented in Table 7. The results indicate

Table 6

The log-likelihood and BIC statistics of SPD BANK for high-frequency in-sample trading data from 1/12/2014 to 29/4/2015 under LGI, Heston, intraday GARCH, intraday EGARCH and ARSV.

	LGI	Heston	GARCH	EGARCH	ARSV
log L	**9542363**	9497876	9374201	9366458	9371597
BIC	**-19084626**	-18995667	-18748317	-18627390	-18743094

Table 7

The one-day-ahead log predictive score (log S) and BIC statistics of SPD BANK for high-frequency out-of-sample data under LGI, Heston, intraday GARCH, intraday EGARCH and ARSV with the sliding window, shown as mean (standard deviation) over the last 20 trading days from 1/4/2015 to 29/4/2015.

	LGI	Heston	GARCH	EGARCH	ARSV
log S	**96290.8**	96274.2	95559.0	95838.8	95733.8
	(1515.8)	(1292.3)	(937.7)	(1053.6)	(959.2)
BIC	**-192514.6**	-192510.2	-191051.0	-191610.6	-191400.6
	(3031.5)	(2584.6)	(1875.4)	(2107.3)	(1918.5)

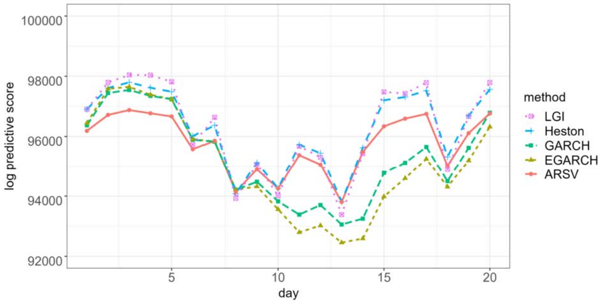


Fig 5. *The one-day-ahead log predictive score for high-frequency out-of-sample data under LGI, Heston, intraday GARCH, intraday EGARCH and ARSV with the sliding window over the last 20 trading days from 1/4/2015 to 29/4/2015.*

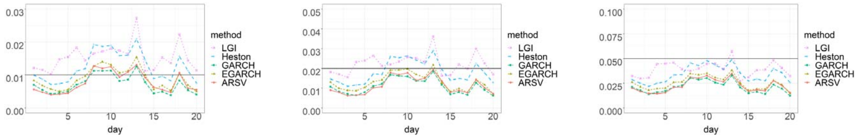
that the LGI model outperforms the other models, achieving the highest predictive likelihood and the lowest BIC. The specific one-day-ahead log predictive score for each sliding window is presented in Figure 5. It can be observed that in most cases, the LGI model exhibits the best performance among the evaluated models. Therefore, we can conclude that the LGI model is also effective for spot volatility estimation and prediction.

6.3. VaR

Value-at-risk (VaR) is a commonly used early-warning indicator in financial risk management. It measures the maximum loss of financial assets over a period at a given confidence level. In this section, we compare the proposed method with its competitors in terms of one-day-ahead forecasting of high-frequency VaR. Given the observations on a sliding window W_d, the predictive high-frequency VaR during $[t_{j-1}, t_j) \subset [d, d+1)$ can be defined as

$$VaR_j^a = u_a \tilde{\sigma}_{t_j} \Delta^{1/2},$$

where u_a is the a-quantile of a standard norm distribution and $\tilde{\sigma}_{t_j}^2$ denotes the forecasting of spot volatility at observation time t_j. VaR_j^a measures the predictive minimum high-frequency per-second log-return of the efficient price at the $1-a$ confidence level. Therefore, if the forecasting is accurate, the probability that the log-return on $[t_{j-1}, t_j)$ falls below VaR_j^a (failure rate) will be exactly a. To evaluate the forecasting of VaR for the models, we consider three cases with $a = 0.01$, 0.02, and 0.05, respectively, and compare the empirical failure rate $\hat{a}^{(d)} := \Delta \sum_{j=md+1}^{(m+1)d} I_{\{\Delta \hat{X}_j < VaR_j^a\}}$ with its target a. Figure 6 shows the empirical failure rates for the LGI model and its competitors over the 20 sliding-window subsets. The competing methods tend to underestimate the VaR, leading to fewer VaR failures than expected at $a = 0.02$ and $a = 0.05$. Therefore, investment strategies based on them will turn out to be too conservative and may miss profitable opportunities compared to our method.



(a) a=0.01 (b) a=0.02 (c) a=0.05

Fig 6. *The empirical failure rate â for SPD BANK under LGI, Heston, intraday GARCH, intraday EGARCH, and ARSV over the 20 sliding windows with the true failure rate* $a = 0.01, 0.02,$ *and* 0.05*, respectively.*

7. Conclusion

This paper proposes a novel Bayesian LGI model for high-frequency data analysis. We assume that the log price of an asset obeys an Itô process and embed the discrete-time GARCH structure into the high-frequency volatility process on a series of latent anchors. By introducing the latent anchors, our model enjoys large flexibility in modeling high-frequency volatility. Furthermore, the latent anchors are approximated with the similar posterior distribution to compute more efficiently. We propose an efficient two-stage Bayesian inference procedure for the anchors as well as the spot and integral volatility, with a corresponding two-stage MCMC sampling algorithm based on the MH and birth-death schemes. In the simulation and empirical study, our method usually yields better results in estimating and predicting volatility compared with its competitors. The latent anchors estimated from real data reveal that the market information is fast accumulated during the opening and closing times. As discussed in Section 4.3, the proposed model can be potentially extended in two directions. Firstly, it could be applied to discontinuous price processes that incorporate jump terms, allowing for the simultaneous modeling and estimation of the volatility processes and jumps. Secondly, the model could facilitate multi-asset analysis, particularly under observations that are not synchronous, thus providing a useful tool for portfolio management. We leave detailed research on these extensions for future studies.

Appendix A: Proofs

A.1. Proof of Proposition 2.1

$$
\begin{aligned}
&\int_{\tau_{i-1}}^{\tau_i} \exp\{\beta(\tau_i - t)\}\sigma_t^2 dt \\
=&\int_{\tau_{i-1}}^{\tau_i} \exp\{\beta(\tau_i - t)\}\{(\tau_i - \tau_{i-1}) - (\tau_i - t)\}w dt \\
+&\int_{\tau_{i-1}}^{\tau_i} \exp\{\beta(\tau_i - t)\}\left(\exp[\gamma\{(\tau_i - \tau_{i-1}) - (\tau_i - t)\}]\right)\sigma_{\tau_{i-1}}^2 dt
\end{aligned}
$$

$$
\begin{aligned}
&+\beta\int_{\tau_{i-1}}^{\tau_i}\int_s^{\tau_i}\exp\{\beta(\tau_i-t)\}\sigma_s^2dtds\\
&+2\beta\int_{\tau_{i-1}}^{\tau_i}\int_s^{\tau_i}\exp\{\beta(\tau_i-t)\}dt(\int_{\tau_{i-1}}^{s}\sigma_h dB_h)\sigma_s dB_s\\
&=\frac{w}{\beta}\left\{-\Delta\tau_i+\frac{\exp(\beta\Delta\tau_i)-1}{\beta}\right\}+\exp\left\{\gamma(\Delta\tau_i)\right\}\sigma_{\tau_{i-1}}^2\frac{\exp\{(\beta-\gamma)\Delta\tau_i\}-1}{\beta-\gamma}\\
&+\int_{\tau_{i-1}}^{\tau_i}\exp\{\beta(\tau_i-t)\}\sigma_t^2dt-\int_{\tau_{i-1}}^{\tau_i}\sigma_t^2dt\\
&+2\int_{\tau_{i-1}}^{\tau_i}[\exp\{\beta(\tau_i-s)\}-1]\int_{\tau_{i-1}}^{t}\sigma_s dB_s\sigma_t dB_t,
\end{aligned}
$$

which means that

$$
\begin{aligned}
\int_{\tau_{i-1}}^{\tau_i}\sigma_t^2dt=&\frac{w}{\beta}\left\{-\Delta\tau_i+\frac{\exp(\beta\Delta\tau_i)-1}{\beta}\right\}\\
&+\exp\left\{\gamma(\Delta\tau_i)\right\}\sigma_{\tau_{i-1}}^2\frac{exp\{(\beta-\gamma)\Delta\tau_i\}-1}{\beta-\gamma}\\
&+2\int_{\tau_{i-1}}^{\tau_i}[\exp\{\beta(\tau_i-s)\}-1]\int_{\tau_{i-1}}^{t}\sigma_s dB_s\sigma_t dB_t,
\end{aligned}
$$

where we denote $D_i=2\int_{\tau_{i-1}}^{\tau_i}[\exp\{\beta(\tau_i-s)\}-1]\int_{\tau_{i-1}}^{t}\sigma_s dB_s\sigma_t dB_t$ as a martiangle difference when given latent anchors τ_i,τ_{i-1},

$$
\begin{aligned}
h_i(\boldsymbol{\xi})&=\frac{w}{\beta}\left\{-\Delta\tau_i+\frac{\exp(\beta\Delta\tau_i)-1}{\beta}\right\}+\frac{\exp\{(\beta-\gamma)\Delta\tau_i\}-1}{\beta-\gamma}\exp(\gamma\Delta\tau_i)\sigma_{\tau_{i-1}}^2\\
&=\frac{w}{\beta}\left\{-\Delta\tau_i+\frac{\exp(\beta\Delta\tau_i)-1}{\beta}\right\}+\frac{\exp(\beta\Delta\tau_i)-\exp(\gamma\Delta\tau_i)}{\beta-\gamma}\sigma_{\tau_{i-1}}^2
\end{aligned}
$$

and can be seen as $h_i(\boldsymbol{\xi})=E_{\boldsymbol{\xi}}(\int_{\tau_{i-1}}^{\tau_i}\sigma_t^2dt|\mathcal{F}_{\tau_{i-1}},\tau_i)$. As we know, the spot volatility at latent anchor $\sigma_{\tau_{i-1}}^2$ can be expanded as:

$$
\begin{aligned}
\sigma_{\tau_{i-1}}^2&=\Delta\tau_{i-1}w+\exp(\gamma\Delta\tau_{i-1})\sigma_{\tau_{i-2}}^2+\beta Z_{\tau_{i-1}}^2(\tau_{i-2})\\
&=w\sum_{l=0}^{\infty}\Delta\tau_{i-1-l}\exp\left(\gamma\sum_{k=1}^{l}\Delta\tau_{i-k}\right)\\
&+\beta\sum_{l=0}^{\infty}\exp\left(\gamma\sum_{k=1}^{l}\Delta\tau_{i-k}\right)Z_{\tau_{i-1-l}}^2(\tau_{i-2-l})\\
&=w\sum_{l=0}^{\infty}\Delta\tau_{i-1-l}\exp\left(\gamma\sum_{k=1}^{l}\Delta\tau_{i-k}\right)+\beta Z_{\tau_{i-1}}^2(\tau_{i-2})\\
&+\beta\sum_{l=1}^{\infty}\exp\left(\gamma\sum_{k=1}^{l}\Delta\tau_{i-k}\right)Z_{\tau_{i-1-l}}^2(\tau_{i-2-l}).
\end{aligned}
$$

We combine the expression of $h_i(\boldsymbol{\xi})$ and $\sigma^2_{\tau_{i-1}}$,

$$\begin{aligned}
&h_i(\boldsymbol{\xi})\\
=&\frac{w}{\beta}\left(-\Delta\tau_i+\frac{\exp(\beta\Delta\tau_i)-1}{\beta}\right)+\frac{\exp(\beta\Delta\tau_i)-\exp(\gamma\Delta\tau_i)}{\beta-\gamma}\\
&\times\{w\sum_{l=0}^{\infty}\Delta\tau_{i-1-l}\exp\left(\gamma\sum_{k=1}^{l}\Delta\tau_{i-k}\right)+\beta Z^2_{\tau_{i-1}}(\tau_{i-2})\\
&+\beta\sum_{l=1}^{\infty}\exp\left(\gamma\sum_{k=1}^{l}\Delta\tau_{i-k}\right)Z^2_{\tau_{i-1-l}}(\tau_{i-2-l})\}.
\end{aligned}$$

With the prior distribution of $\Delta\tau_i\sim Exp(1/\alpha), i=1,\cdots,p$ and the independence of $\Delta\tau_i, i=1,\cdots,p$, we deduce that $E_{\boldsymbol{\xi}}(\Delta\tau_i)=\alpha$, $E_{\boldsymbol{\xi}}(e^{\beta\Delta\tau_i})=1/(1-\alpha\beta)$. As a result,

$$\begin{aligned}
&E_{\boldsymbol{\xi}}\{h_i(\boldsymbol{\xi})\}\\
=&\frac{\alpha^2 w}{1-\alpha\beta}+\frac{\frac{1}{1-\alpha\beta}-\frac{1}{1-\alpha\gamma}}{\beta-\gamma}\\
&\times\left\{-\frac{w(1-\alpha\gamma)}{\gamma}+\beta Z^2_{\tau_{i-1}}(\tau_{i-2})+\beta\sum_{l=1}^{\infty}\left(\frac{1}{1-\alpha\gamma}\right)^l Z^2_{\tau_{i-1-l}}(\tau_{i-2-l})\right\}\\
=&\frac{-\alpha^2 w\alpha\gamma+w\alpha^2}{(1-\alpha\beta)(1-\alpha\gamma)}+\frac{\alpha\beta}{(1-\alpha\beta)(1-\alpha\gamma)}Z^2_{\tau_{i-1}}(\tau_{i-2})+\frac{1}{1-\alpha\gamma}E_{\boldsymbol{\xi}}\{h_{i-1}(\boldsymbol{\xi})\}\\
=&\frac{\alpha^2 w}{1-\alpha\beta}+\frac{\alpha\beta}{(1-\alpha\beta)(1-\alpha\gamma)}Z^2_{\tau_{i-1}}(\tau_{i-2})+\frac{1}{1-\alpha\gamma}E_{\boldsymbol{\xi}}\{h_{i-1}(\boldsymbol{\xi})\},
\end{aligned}$$

which means that,

$$E_{\boldsymbol{\xi}}\left(\int_{\tau_{i-1}}^{\tau_i}\sigma_t^2 dt|\mathcal{F}_{\tau_{i-1}}\right)=w^g+\gamma^g E_{\boldsymbol{\xi}}\left(\int_{\tau_{i-2}}^{\tau_{i-1}}\sigma_t^2 dt|\mathcal{F}_{\tau_{i-2}}\right)+\beta^g Z^2_{\tau_{i-1}}(\tau_{i-2}).$$

where $w^g=w\alpha^2/(1-\alpha\beta)$, $\gamma^g=1/(1-\alpha\gamma)$, $\beta^g=\alpha\beta/\{(1-\alpha\beta)(1-\alpha\gamma)\}$.

A.2. Proof of Theorem 3.1

Lemma A.1. *Under Assumption 2.1,*

$$E_{\boldsymbol{\theta}}(\sigma^2_{t_j})\le C,\quad E_{\boldsymbol{\theta}}(\tilde{\sigma}^2_{t_j})\le C,$$

for every $j=1,\cdots,n$.

Proof. We proof this conclusion with the induction method. Firstly, for $t_j\le\tau_1$

$$\begin{aligned}
E_{\boldsymbol{\theta}}(\sigma^2_{t_j})&=t_j w+\exp(\gamma t_j)\sigma_0^2+E_{\boldsymbol{\theta}}\left(\int_0^{t_j}\sigma_s^2 ds\right)\\
&\le t_j w+\exp(\gamma t_j)\sigma_0^2+E_{\boldsymbol{\theta}}\left(\int_0^{\tau_j}\sigma_s^2 ds\right)\\
&\le C.
\end{aligned}$$

The last inequality satisfies for that the integral process $\int_{\tau_{i-1}}^{\tau_i}\sigma_s^2 ds$ is stationary under Proposition 2.1 with $E_{\boldsymbol{\theta}}(\int_0^{\tau_1}\sigma_s^2 ds) = C_1$, $C, C_1 > 0$ are constant and C is far greater than C_1.

Then, if $E(\sigma_{t_j}^2) \leq C$ with $t_j \in (\tau_{i-1}, \tau_i]$, we can deduce that

$$\begin{aligned}
E_{\boldsymbol{\theta}}(\sigma_{t_j}^2) &= (t_j - \tau_i)w + \exp\{\gamma(t_j - \tau_i)\}E_{\boldsymbol{\theta}}(\sigma_{\tau_i}^2) + E_{\boldsymbol{\theta}}\left(\int_{\tau_i}^{t_j}\sigma_s^2 ds\right)\\
&\leq (t_j - \tau_i)w + \exp\{\gamma(t_j - \tau_i)\}E_{\boldsymbol{\theta}}(\sigma_{\tau_i}^2) + E_{\boldsymbol{\theta}}\left(\int_{\tau_i}^{\tau_{i+1}}\sigma_s^2 ds\right)\\
&\leq (t_j - \tau_i)w + C\exp\{\gamma(t_j - \tau_i)\} + C_1\\
&\leq Tw + C\exp\{\gamma\Delta t_j\} + C_1\\
&\leq C.
\end{aligned}$$

There must exists the constant C to satisfy the last inequality for that $\exp\{\gamma(t_j - \tau_i)\} < 1$.

Therefore, we can say that for every $j = 1, \cdots, n$,

$$E_{\boldsymbol{\theta}}(\sigma_{t_j}^2) \leq C.$$

Similarly, we can proof that $E_{\boldsymbol{\theta}}(\tilde{\sigma}_{t_j}^2) \leq C$. □

Proof of Theorem 3.1: First, We rewrite $\sigma_{t_j}^2 = f(\boldsymbol{\tau}, X_{\boldsymbol{\tau}}, t_j, X_{t_j}, \boldsymbol{\xi}), \tilde{\sigma}_{t_j}^2 = f(\tilde{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}}, t_j, X_{t_j}, \boldsymbol{\xi})$, where $f(\cdot)$ is a function. Besides, $\pi(\boldsymbol{\theta}|\mathbf{D}), \tilde{\pi}(\tilde{\boldsymbol{\theta}}|\mathbf{D})$ are the function of $(\boldsymbol{\sigma}^2, \mathbf{X}), (\tilde{\boldsymbol{\sigma}}^2, \mathbf{X})$ respectively, where $\boldsymbol{\sigma}^2 = (\sigma_{t_1}^2, \cdots, \sigma_{t_n}^2)$, $\tilde{\boldsymbol{\sigma}}^2 = (\tilde{\sigma}_{t_1}^2, \cdots, \tilde{\sigma}_{t_n}^2)$. To simplify, we denote

$$b(\boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}, X_{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}}) = \pi(\mathbf{D}|\tilde{\boldsymbol{\theta}}) - \pi(\mathbf{D}|\boldsymbol{\theta})$$

with $c(\boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}, X_{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}}) = \pi(\tilde{\boldsymbol{\theta}}|\mathbf{D}) - \int_{\boldsymbol{\tau}\in\mathcal{G}}\pi(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\tau}$. Beside, we denote $\tilde{c} = \int\pi(\mathbf{D}|\tilde{\boldsymbol{\theta}})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, $c = \int\pi(\mathbf{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. Therefore, $\pi(\boldsymbol{\theta}|\mathbf{D}) = \pi(\mathbf{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/c$, $\tilde{\pi}(\tilde{\boldsymbol{\theta}}|\mathbf{D}) = \pi(\mathbf{D}|\tilde{\boldsymbol{\theta}})\pi(\boldsymbol{\theta})/\tilde{c}$.

$$\begin{aligned}
&c(\boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}, X_{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}})\\
&= \frac{\pi(\mathbf{D}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})}{\tilde{c}} - \int_{\boldsymbol{\tau}\in\mathcal{G}}\pi(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\tau}\\
&\approx \frac{\pi(\mathbf{D}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})}{\tilde{c}} - \pi(\boldsymbol{\theta}|\mathbf{D})\prod_{i=1}^{p}\Delta t_{j_i}\\
&= \frac{b(\boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}, X_{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}})\pi(\tilde{\boldsymbol{\theta}})}{\tilde{c}} + \frac{\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\pi(\tilde{\boldsymbol{\theta}})}{\tilde{c}} - \pi\left(\boldsymbol{\theta}|\mathbf{D}\right)\prod_{i=1}^{p}\Delta t_{j_i}\\
&= \frac{b(\boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}, X_{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}})\pi(\tilde{\boldsymbol{\theta}})}{\tilde{c}}\\
&+ \pi\left(\boldsymbol{\theta}|\mathbf{D}\right)\left(\frac{c\{\pi(\tilde{\boldsymbol{\theta}}) - \pi(\boldsymbol{\theta})\prod_{i=1}^{p}\Delta t_{j_i}\} + \pi(\boldsymbol{\theta})\prod_{i=1}^{p}\Delta t_{j_i}(c - \tilde{c})}{\tilde{c}\pi(\boldsymbol{\theta})}\right).
\end{aligned}$$

Then

$$\begin{aligned}\frac{c(\boldsymbol{\tau},\tilde{\boldsymbol{\tau}},X_{\boldsymbol{\tau}},X_{\tilde{\boldsymbol{\tau}}})}{\int_{\boldsymbol{\tau}\in\mathcal{G}}\pi(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\tau}} \approx{}& \frac{b(\boldsymbol{\tau},\tilde{\boldsymbol{\tau}},X_{\boldsymbol{\tau}},X_{\tilde{\boldsymbol{\tau}}})}{\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}\frac{c\pi(\tilde{\boldsymbol{\theta}})}{\tilde{c}\pi(\boldsymbol{\theta})\prod_{i=1}^{p}\Delta t_{j_i}}\\ &+\frac{c\{\pi(\tilde{\boldsymbol{\theta}})-\pi(\boldsymbol{\theta})\prod_{i=1}^{p}\Delta t_{j_i}\}+\pi(\boldsymbol{\theta})\prod_{i=1}^{p}\Delta t_{j_i}(c-\tilde{c})}{\tilde{c}\pi(\boldsymbol{\theta})\prod_{i=1}^{p}\Delta t_{j_i}}.\end{aligned}$$

We can deduce $b(\boldsymbol{\tau},\tilde{\boldsymbol{\tau}},X_{\boldsymbol{\tau}},X_{\tilde{\boldsymbol{\tau}}})$ with the Talyer expansion

$$\begin{aligned}b(\boldsymbol{\tau},\tilde{\boldsymbol{\tau}},X_{\boldsymbol{\tau}},X_{\tilde{\boldsymbol{\tau}}})=\sum_{i=1}^{p}\Bigg\{&\frac{\partial\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{\partial\tau_i}(\tilde{\tau}_i-\tau_i)+\frac{\partial\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{\partial X_{\tau_i}}(X_{\tilde{\tau}_i}-X_{\tau_i})\\ &+\frac{1}{2}\frac{\partial^2\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{(\partial X_{\tau_i})^2}(X_{\tilde{\tau}_i}-X_{\tau_i})^2+o(\tilde{\tau}_i-\tau_i)+o(X_{\tilde{\tau}_i}-X_{\tau_i})^2\Bigg\}.\end{aligned}$$

As we know,

$$\frac{\partial\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{\partial\sigma^2_{t_j}}=\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}},\qquad\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}=\exp\{\gamma(t_{j-1}-\tau_i)\}.$$

Additionally, $\sigma^2_{t_j}$ is independent with τ_i and X_{τ_i} when $t_j<\tau_i$. Therefore, for every integer $i\le p$, the deviation can be deduced as

$$\begin{aligned}\frac{\partial\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{\partial\tau_i}&=\sum_{j=1}^{n}\frac{\partial\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{\partial\sigma^2_{t_j}}\frac{\partial\sigma^2_{t_j}}{\tau_i}\\ &=\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\sum_{j=\tau_i+2}^{\tau_{i+1}+1}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}\left[-w-\gamma\exp\{\gamma(t_{j-1}-\tau_i)\}\sigma^2_{\tau_i}\right]\\ &\quad+\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\sum_{j=\tau_i+1}^{n}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}\left[w+\gamma\exp\{\gamma(\tau_i-\tau_{i-1})\}\sigma^2_{\tau_{i-1}}\right],\\ \frac{\partial\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{\partial X_{\tau_i}}&=\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\sum_{j=\tau_i+2}^{\tau_{i+1}+1}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}\left\{-2\beta(X_{t_{j-1}}-X_{\tau_i})\right\}\\ &\quad+\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\sum_{j=\tau_i+1}^{n}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}\left\{2\beta(X_{\tau_i}-X_{\tau_{i-1}})\right\},\\ \frac{\partial^2\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{(\partial X_{\tau_i})^2}&=\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\Bigg(\sum_{j=\tau_i+2}^{\tau_{i+1}+1}\left\{-2\beta(X_{t_{j-1}}-X_{\tau_i})\right\}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}\\ &\quad+\sum_{j=\tau_i+1}^{n}\left\{2\beta(X_{\tau_i}-X_{\tau_{i-1}})\right\}\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}\Bigg)^2\end{aligned}$$

$$
\begin{aligned}
&+\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)2\beta\left\{\sum_{j=\tau_i+2}^{\tau_{i+1}+1}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}+\sum_{j=\tau_i+1}^{n}\frac{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1}{2\sigma^2_{t_{j-1}}}\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}\right\}\\
&+\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\sum_{j=\tau_i+2}^{\tau_{i+1}+1}\left\{-2\beta(X_{t_{j-1}}-X_{\tau_i})\right\}^2\left\{\frac{2-4(\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}})}{4(\sigma^2_{t_{j-1}})^2}\right\}\\
&+\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)\sum_{j=\tau_i+1}^{n}\left\{2\beta(X_{\tau_i}-X_{\tau_{i-1}})\right\}^2\left\{\frac{2-4(\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}})}{4(\sigma^2_{t_{j-1}})^2}\right\}\left(\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}\right)^2
\end{aligned}
$$

For that $(\Delta X_{t_j})^2/(\Delta t_j\sigma^2_{t_{j-1}})\sim\chi^2(1)$ and is independent with $\sigma^2_{t_j}$, $E_{\boldsymbol{\theta}}\{(\Delta X_{t_j})^2/(\Delta t_j\sigma^2_{t_{j-1}})\}=1$, $E_{\boldsymbol{\theta}}\{(\Delta X_{t_j})^2/(\Delta t_j\sigma^2_{t_{j-1}})\}^2=3$. So that

$$
\begin{aligned}
&E_{\boldsymbol{\theta}}\left\{\frac{1}{\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}\frac{\partial^2\pi\left(\mathbf{D}|\boldsymbol{\theta}\right)}{(\partial X_{\tau_i})^2}\right\}\\
&=\sum_{j=\tau_i+2}^{\tau_{i+1}+1}E_{\boldsymbol{\theta}}\frac{[-2\beta(X_{t_{j-1}}-X_{\tau_i})]^2}{(2\sigma^2_{t_{j-1}})^2}E_{\boldsymbol{\theta}}\left\{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1\right\}^2\\
&+\sum_{j=\tau_i+1}^{n}E_{\boldsymbol{\theta}}\frac{\{2\beta(X_{\tau_i}-X_{\tau_{i-1}})\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}\}^2}{(2\sigma^2_{t_{j-1}})^2}E_{\boldsymbol{\theta}}\left\{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1\right\}^2\\
&+2\beta\left[\sum_{j=\tau_i+2}^{\tau_{i+1}+1}E_{\boldsymbol{\theta}}\left\{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1\right\}E_{\boldsymbol{\theta}}\frac{1}{2\sigma^2_{t_{j-1}}}\right.\\
&\left.+\sum_{j=\tau_i+1}^{n}\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}E_{\boldsymbol{\theta}}\left\{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1\right\}E_{\boldsymbol{\theta}}\frac{1}{2\sigma^2_{t_{j-1}}}\right]\\
&+\sum_{j=\tau_i+2}^{\tau_{i+1}+1}E_{\boldsymbol{\theta}}\frac{\{-2\beta(X_{t_{j-1}}-X_{\tau_i})\}^2}{4(\sigma^2_{t_{j-1}})^2}E_{\boldsymbol{\theta}}\left\{2-4\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}\right\}\\
&+\sum_{j=\tau_i+1}^{n}E_{\boldsymbol{\theta}}\frac{\{2\beta(X_{\tau_i}-X_{\tau_{i-1}})\}^2(\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}})^2}{4(\sigma^2_{t_{j-1}})^2}E_{\boldsymbol{\theta}}\left\{2-4\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}\right\}\\
&=\sum_{j=\tau_i+2}^{\tau_{i+1}+1}E_{\boldsymbol{\theta}}\frac{\{-2\beta(X_{t_{j-1}}-X_{\tau_i})\}^2}{(2\sigma^2_{t_{j-1}})^2}\\
&\times\left[E_{\boldsymbol{\theta}}\left\{\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}-1\right\}^2+E_{\boldsymbol{\theta}}\left\{2-4\frac{(\Delta X_{t_j})^2}{\Delta t_j\sigma^2_{t_{j-1}}}\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{j=\tau_i+1}^{n} E_{\boldsymbol{\theta}} \frac{\{2\beta(X_{\tau_i} - X_{\tau_{i-1}})\frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}}\}^2}{(2\sigma^2_{t_{j-1}})^2} \\
&\times \left[E_{\boldsymbol{\theta}} \left\{ \frac{(\Delta X_{t_j})^2}{\Delta t_j \sigma^2_{t_{j-1}}} - 1 \right\}^2 + E_{\boldsymbol{\theta}} \left\{ 2 - 4\frac{(\Delta X_{t_j})^2}{\Delta t_j \sigma^2_{t_{j-1}}} \right\} \right] \\
&+ 2\beta \left[\sum_{j=\tau_i+2}^{\tau_{i+1}+1} E_{\boldsymbol{\theta}} \left\{ \frac{(\Delta X_{t_j})^2}{\Delta t_j \sigma^2_{t_{j-1}}} - 1 \right\} E_{\boldsymbol{\theta}} \frac{1}{2\sigma^2_{t_{j-1}}} \right. \\
&\left. + \sum_{j=\tau_i+1}^{n} \frac{\partial\sigma^2_{t_{j-1}}}{\partial\sigma^2_{\tau_i}} E_{\boldsymbol{\theta}} \left\{ \frac{(\Delta X_{t_j})^2}{\Delta t_j \sigma^2_{t_{j-1}}} - 1 \right\} E_{\boldsymbol{\theta}} \frac{1}{2\sigma^2_{t_{j-1}}} \right] \\
&= 0,
\end{aligned}
$$

The last equation satisfies for that $E_{\boldsymbol{\theta}}\{2 - 4(\Delta X_{t_j})^2/(\Delta t_j \sigma^2_{t_{j-1}})\} = -2$ and $E_{\boldsymbol{\theta}}\{(\Delta X_{t_j})^2/(\Delta t_j \sigma^2_{t_{j-1}}) - 1\}^2 = 2$. Besides,

$$
E_{\boldsymbol{\theta}} \left\{ \frac{1}{\pi(\mathbf{D}|\boldsymbol{\theta})} \frac{\partial \pi(\mathbf{D}|\boldsymbol{\theta})}{\partial \tau_i} \right\} = E_{\boldsymbol{\theta}} \left\{ \frac{1}{\pi(\mathbf{D}|\boldsymbol{\theta})} \frac{\partial \pi(\mathbf{D}|\boldsymbol{\theta})}{\partial X_{\tau_i}} \right\} = 0.
$$

Therefore,

$$
\begin{aligned}
E_{\boldsymbol{\theta}} \left\{ \frac{b(\boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}, X_{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}})}{\pi(\mathbf{D}|\boldsymbol{\theta})} \right\} &= \sum_{i=1}^{p} o(\tau_i - \tilde{\tau}_i) + o(E_{\boldsymbol{\theta}}(X_{\tau_i} - X_{\tilde{\tau}_i})^2) \\
&= o(n^{\nu_1 - 1}) + \sum_{i=1}^{p} E_{\boldsymbol{\theta}}(\sigma^2_{\tilde{\tau}_i}) o(\tau_i - \tilde{\tau}_i) \\
&= o(n^{\nu_1 - 1}).
\end{aligned}
$$

The last equation satisfies with $E_{\boldsymbol{\theta}}(\sigma^2_{\tilde{\tau}_i}) \leq C$ under Lemma A.1 and $p \leq n^{\nu_1}$ under Assumption 3.2. On the other hand,

$$
\pi(\tilde{\boldsymbol{\tau}}|p, \alpha) = \alpha^{-p} exp(-\frac{T}{\alpha}) I_{\{0 < t_{j_1} < \cdots < t_{j_p} \leq T\}} \prod_{i=1}^{p} \Delta t_{j_i},
$$

As that the order of the latent anchors will not change after the approximation scheme, we can deduce that $\pi(\tilde{\boldsymbol{\theta}}) - \pi(\boldsymbol{\theta}) \prod_{i=1}^{p} \Delta t_{j_i} = 0$. Also, $E_{\boldsymbol{\theta}}(c/\tilde{c}) = O(1)$, $E_{\boldsymbol{\theta}}((c - \tilde{c})/\tilde{c}) = o(n^{\nu_1 - 1})$. So that

$$
E_{\boldsymbol{\theta}} \left\{ \frac{c(\boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}, X_{\boldsymbol{\tau}}, X_{\tilde{\boldsymbol{\tau}}})}{\int_{\boldsymbol{\tau} \in \mathcal{G}} \pi(\boldsymbol{\theta}|\mathbf{D}) \, d\boldsymbol{\tau}} \right\} = o(n^{\nu_1 - 1}).
$$

Appendix B: Detail of MCMC algorithm

The specified MCMC algorithm based on Gibbs sampling can be decomposed to 4 steps:

Update market microstructure noise $\boldsymbol{\delta}$

$$\pi(\boldsymbol{\delta}|\boldsymbol{\xi}, p, \tilde{\boldsymbol{\tau}}, \sigma_0^2, \mathbf{D}) \propto \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\left\{-\frac{(\Delta X_{t_j})^2}{2\sigma_{t_{j-1}}^2 \Delta t_j}\right\},$$

where $\Delta X_{t_j} = X_{t_j} - X_{t_{j-1}}$, $X_{t_j} = Y_{t_j} - h(Z_{t_j}; \boldsymbol{\delta})$. We use Metropolis Hastings algorithm (MH) to sample $\boldsymbol{\delta}$.

Update latent anchors $(p, \tilde{\boldsymbol{\tau}})$

$$\begin{aligned}
&\pi(p, \tilde{\boldsymbol{\tau}}|\boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2, \mathbf{D}) \\
&\propto \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\left\{-\frac{(\Delta X_{t_j})^2}{2\sigma_{t_{j-1}}^2 \Delta t_j}\right\} \alpha^{-p} e^{-\frac{T}{\alpha}} I_{0<\tau_1<\cdots<\tau_p \leq T}, \\
&\pi(\tilde{\tau}_i|p, \tilde{\boldsymbol{\tau}}_{(-i)}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma_0^2, \mathbf{D}) \propto \prod_{\tilde{\tau}_{i-1}<t_j<\tilde{\tau}_{i+1}} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\left\{-\frac{(\Delta X_{t_j})^2}{2\sigma_{t_{j-1}}^2 \Delta t_j}\right\},
\end{aligned}$$

where $\tilde{\boldsymbol{\tau}}_{(-i)} = (\tilde{\tau}_1, \cdots, \tilde{\tau}_{i-1}, \tilde{\tau}_{i+1}, \cdots, \tilde{\tau}_p)$. We show the sample steps in Algorithm 1.

Update model parameters $w, \gamma, \beta, \alpha, \sigma_0^2$

$$\pi(w|\gamma, \beta, \alpha, \boldsymbol{\delta}, p, \tilde{\boldsymbol{\tau}}, \sigma_0^2, \mathbf{D}) \propto \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\left\{-\frac{(\Delta X_{t_j})^2}{2\sigma_{t_{j-1}}^2 \Delta t_j}\right\} I_{\{w>0\}}.$$

$$\begin{aligned}
&\pi(\gamma|w, \beta, \alpha, \boldsymbol{\delta}, p, \tilde{\boldsymbol{\tau}}, \sigma_0^2, \mathbf{D}) \\
&\propto \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\left\{-\frac{(\Delta X_{t_j})^2}{2\sigma_{t_{j-1}}^2 \Delta t_j}\right\} I_{\{\gamma<0,(1-\alpha\beta)(1-\alpha\gamma)>1\}}. \\
&\pi(\beta|w, \gamma, \alpha, \boldsymbol{\delta}, p, \tilde{\boldsymbol{\tau}}, \sigma_0^2, \mathbf{D}) \\
&\propto \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\left\{-\frac{(\Delta X_{t_j})^2}{2\sigma_{t_{j-1}}^2 \Delta t_j}\right\} I_{\{\beta>0,(1-\alpha\beta)(1-\alpha\gamma)>1\}}.
\end{aligned}$$

$$\pi(\alpha|w, \gamma, \beta, \boldsymbol{\delta}, p, \tilde{\boldsymbol{\tau}}, \sigma_0^2, \mathbf{D}) \propto \prod_{i=1}^{p} \frac{1}{\alpha} \exp\left\{-\frac{\Delta\tau_i}{a}\right\} I_{\{\alpha>0,(1-\alpha\beta)(1-\alpha\gamma)>1\}}.$$

$$\pi(\sigma_0^2|w, \gamma, \beta, \alpha, \boldsymbol{\delta}, p, \tilde{\boldsymbol{\tau}}, \mathbf{D}) \propto \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{t_{j-1}}^2 \Delta t_j}} \exp\{-\frac{(\Delta X_{t_j})^2}{2\sigma_{t_{j-1}}^2 \Delta t_j}\} I_{\{\sigma_0^2>0\}},$$

where $\sigma_{t_j}^2 = (t_j - t_{\tilde{j}_{i-1}})w + \exp\{\gamma(t_j - t_{\tilde{j}_{i-1}})\}\sigma_{t_{\tilde{j}_{i-1}}}^2 + \beta(X_{t_j} - X_{t_{\tilde{j}_{i-1}}})^2$, $t_j \in (t_{\tilde{j}_{i-1}}, t_{\tilde{j}_i}]$. We use MH algorithm to sample parameters $w, \gamma, \beta, \alpha, \sigma_0^2$ respectively.

Appendix C: Additional simulation results

C.1. Convergence diagnostics

In this subsection, we first simulate a set of data under the Heston model with the parameter setting as SL when the observed frequency $\Delta = 1/500$. Then, we realize the LGI model with MCMC algorithm under 10 different initial values for the parameters respectively. Finally, we applied the Gelman-Rubin diagnostic under the 10 MCMC chains, and the results are shown in Table 8. Note that the test value is less than 1.1, the LGI model is insensitive to the choice of the initial value.

Table 8

The multivariate potential scale reduction factor (multivariate psrf) under Gelman-Rubin diagnostic with 10 different initial values of the parameters for the MCMC chains generated under Scenario 1.

	α	w	γ	β	δ_1	δ_2	σ_0	Multivariate psrf
Point est.	1.02	1.04	1.03	1.04	1.00	1.00	1.00	1.09
Upper C.I.	1.04	1.07	1.07	1.09	1.00	1.01	1.00	

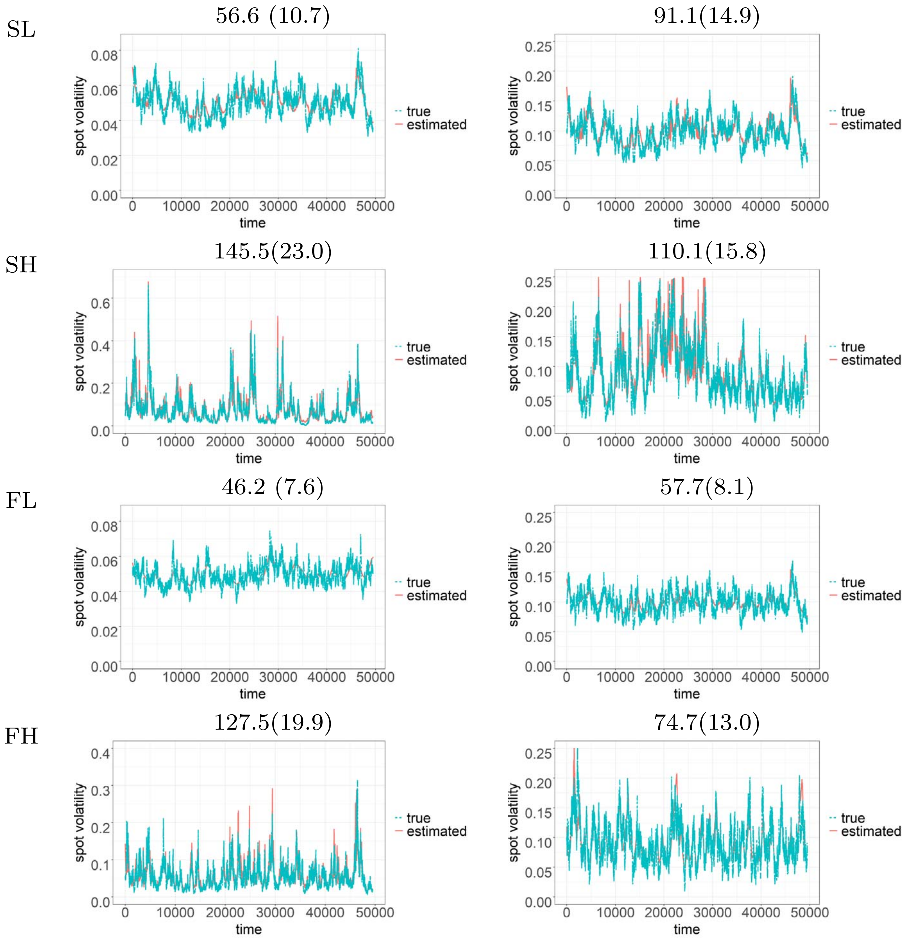
C.2. Estimation of spot volatility with observed frequency $\Delta = 1/500$, $\Delta = 1/5000$

Figure 7, 8 show the true volatility and volatility estimated by LGI model when $\Delta = 1/500, 1/5000$ under 2 scenarios with all settings for a single run respectively. The result is similar to the case of $\Delta = 1/1500$ in the text simulation.

Appendix D: Additional empirical study results

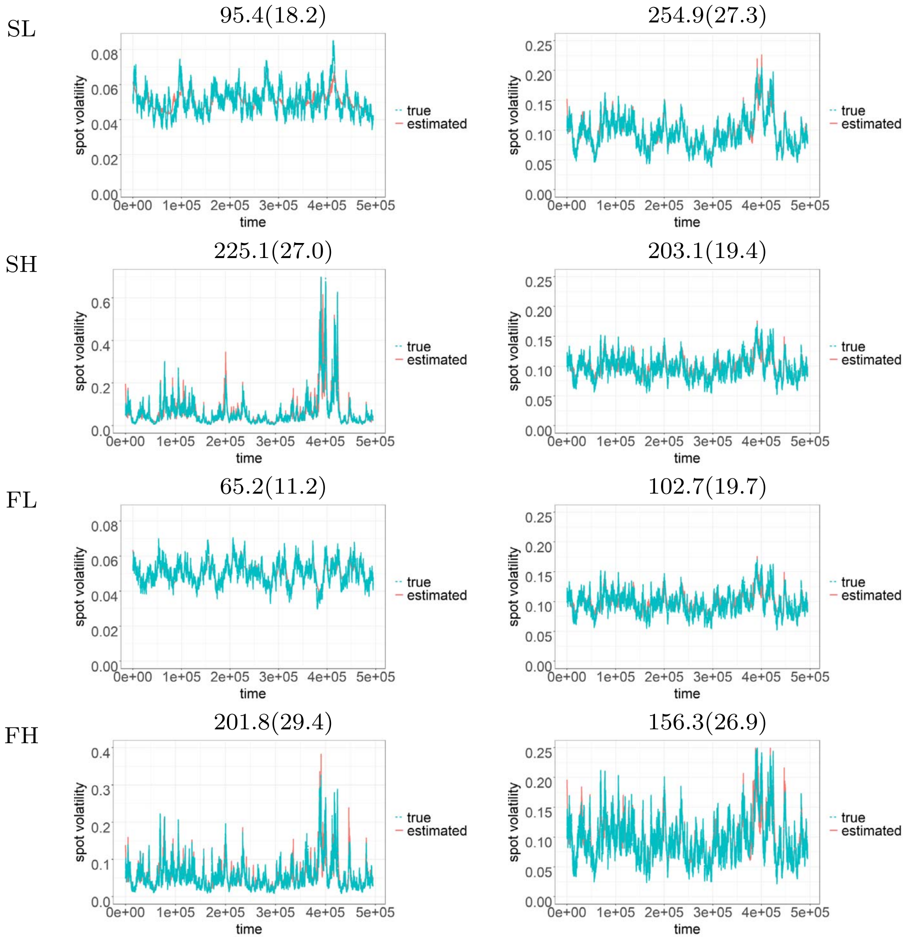
D.1. Posterior convergence analysis and estimation of parameters for SPD BANK

The specific sample path of parameters under the GARCH-Itô model for SPD BANK are shown in Figure 9. The MCMC chains for parameters have been converged under the first 5,000 draws, which means we can draw fewer samples from the Markov chain compared to the setting in Subsection 6.2. Note that the effects of the conditional variance in the previous period γ^g and the squared of log return β^g are $1/(1-\gamma\alpha) = 0.557$, $\alpha\beta/(1-\alpha\beta)(1-\alpha\gamma) = 0.433$ respectively, indicating that their contributions to the conditional variance of low-frequency log return $E_{\boldsymbol{\theta}}[Z^2_{\tau_i}(\tau_{i-1})|\mathcal{F}_{\tau_{i-1}}]$ are relatively similar. Also, the estimation of parameters for trading type and trading volume δ_1, δ_2 are $1.638 \times 10^{-4}, 4.208 \times 10^{-14}$ under LGI model respectively, while are $1.638 \times 10^{-4}, 4.607 \times 10^{-14}$ under ERV method, which indicates that LGI model is comparable with ERV method in the estimation of market microstructure noise.



(a) Scenario 1 (b) Scenario 2

Fig 7. *True volatility and volatility estimated by the LGI model with* $\Delta = 1/500$. *Figure(a) and Figure(b) shows the single-run results under Scenario 1 and Scenario 2 respectively. The mean (standard deviation) of the number of latent anchors over 100 simulation runs are shown on the top of each plot.*



95.4(18.2)

254.9(27.3)

225.1(27.0)

203.1(19.4)

65.2(11.2)

102.7(19.7)

201.8(29.4)

156.3(26.9)

(a) Scenario 1

(b) Scenario 2

FIG 8. *True volatility and volatility estimated by the LGI model with* $\Delta = 1/5000$*. Figure(a) and Figure(b) shows the single-run results under Scenario 1 and Scenario 2 respectively. The mean (standard deviation) of the number of latent anchors over 100 simulation runs are shown on the top of each plot.*

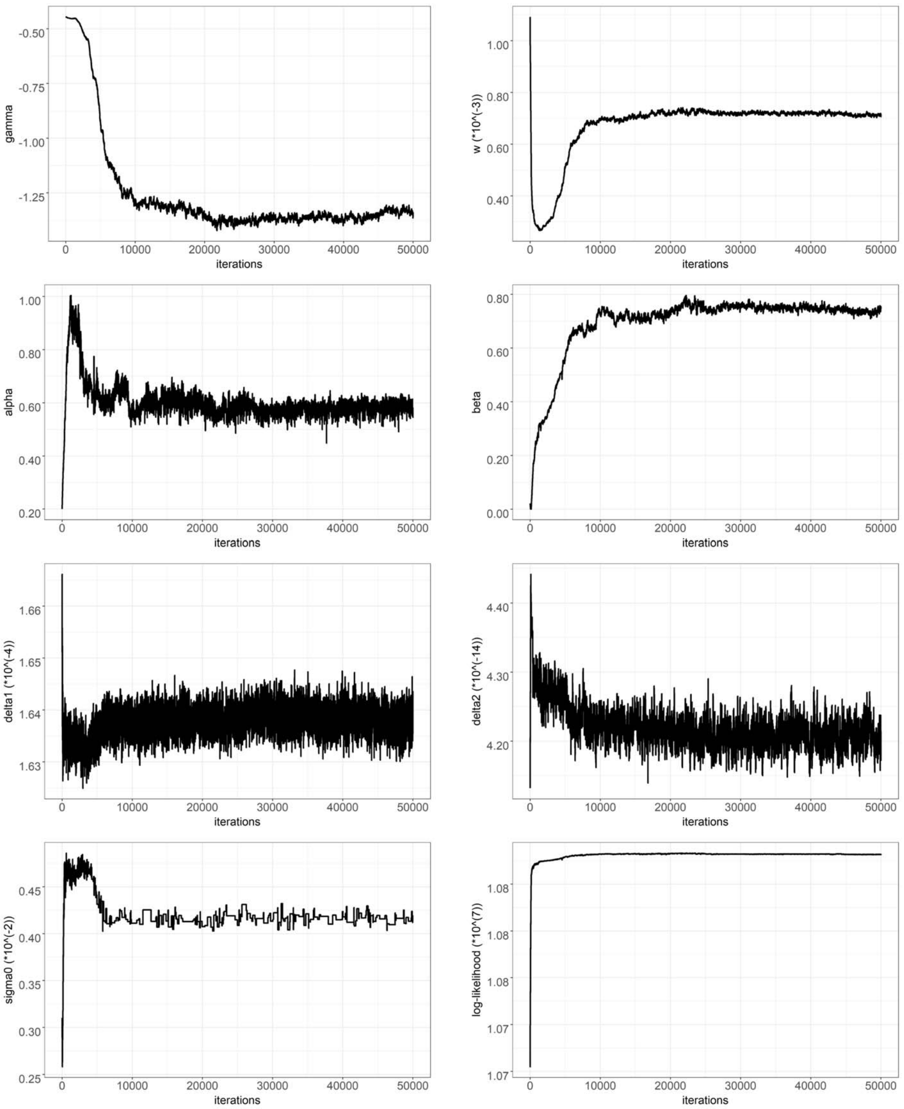


Fig 9. *Sample paths of the Markov chains for the parameters and loglikelihood for SPD BANK.*

TABLE 9

The prediction error of the integral volatility under LGI, GARCH-Itô and ERV with the sliding window, shown as MSE and MAE over the last 20 trading days from 1/4/2015 to 29/4/2015 for stocks CAN and CMBC.

	CAN		CMBC	
	MSE($\times 10^{-4}$)	MAE($\times 10^{-4}$)	MSE($\times 10^{-4}$)	MAE($\times 10^{-4}$)
LGI	**4.977**	**3.795**	**7.512**	**4.765**
GARCH-Itô	5.742	4.460	8.526	5.331
ERV	5.211	4.021	11.687	7.793

TABLE 10

The estimation log-likelihood and BIC statistics for in-sample trading data from 1/12/2014 to 29/4/2015 and one-day-ahead prediction log-likelihood and BIC statistics for high-frequency out-of-sample trading data under LGI, Heston, intraday GARCH, intraday EGARCH and ARSV with the sliding window over the last 20 trading days for stocks CAN and CMBC.

	CAN		CMBC	
Estimation	log likelihood	BIC	log likelihood	BIC
LGI	**9631531**	-19262963	**9098650**	-18197200
Heston	9535752	-19071461	8828705	-17657367
GARCH	9262626	-18525152	8990105	-17980124
EGARCH	9254533	-18508966	8977887	-17955688
ARSV	9554400	-191087025	9060596	-18121093
Prediction	log likelihood	BIC	log likelihood	BIC
LGI	**97715.4**	-195363.8	**91666.0**	-183264.9
	(479.6)	(959.2)	(2521.9)	(5043.7)
Heston	95348.7	-190659.1	91618.7	-183199.0
	(2941.9)	(5883.8)	(2246.6)	(4493.1)
GARCH	91801.5	-183536.1	90818.8	-181580.2
	(944.6)	(1889.3)	(1307.2)	(3327.6)
EGARCH	90848.2	-181629.3	90205.1	-180352.6
	(683.2)	(1366.3)	(1217.3)	(2434.7)
ARSV	95335.3	-190603.6	91379.2	-182700.9
	(2341.1)	(4682.3)	(2009.5)	(4019.0)

D.2. Results of other two stocks

We also take Guangzhou Baiyun International Airport (CAN) and China Minsheng Banking (CMBC) as new case studies. With the same data selection and processing procedure as in Section 6, we get the high-frequency in-sample fitting comparison, high-frequency out-of-sample prediction comparison, low-frequency integral volatility prediction comparison, high-frequency VaR prediction comparison used in risk early warning, and are shown in Table 9, 10 and Figure 10 respectively. We find the same empirical results as in Section 6.

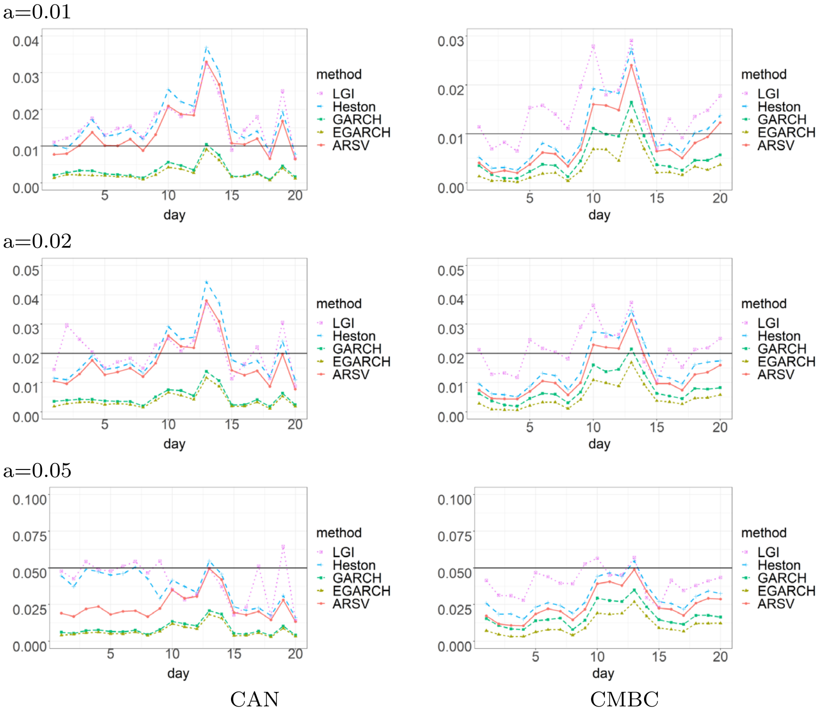


Fig 10. *The empirical failure rate $\hat{a}$ for CAN and CMBC under LGI, Heston, intraday GARCH, intraday EGARCH, and ARSV over the 20 sliding windows with the true failure rate* $a = 0.01, 0.02$, *and* 0.05, *respectively.*

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

Wu's research was supported by National Natural Science Foundation of China (NSFC, 12201628) and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001). Zhang's research was supported by the National natural Science Foundation of China(72271232) and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001). This research was supported by Public Computing Cloud, Renmin University of China.

References

[1] Aït-Sahalia, Y., Fan, J. and Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association* **105** 1504-1517. MR2796567

[2] Aït-Sahalia, Y., Jacod, J. and Li, J. (2010). Testing for jumps in noisy high frequency data. *Journal of Econometrics* **168** 207-222. MR2923764

[3] Almgren, R. and Chriss, N. (2000). Optimal execution of portfolio transactions. *Journal of Risk* **3** 5-39.

[4] Andersen, T. G. and Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* **4** 115-158.

[5] Andersen, T. G. and Bollerslev, T. (1998). DM-Dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer-run dependencies. *Journal of Finance* **53** 219-265.

[6] Andersen, T. G., Bollerslev, T. and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics* **89** 701-720.

[7] Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* **71** 579-625. MR1958138

[8] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2011). Multivariate realised kernels: Consistent positive semidefinite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* **162** 149-169. MR2795610

[9] Barndorff-Nielsen, O. E. and Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* **4** 1-30. MR2051439

[10] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Economics* **31** 307-327. MR0853051

[11] Bollerslev, T., Patton, A. J. and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics* **192** 1-18. MR3463661

[12] Chi, M. (2013). Private information, overconfidence and intraday trading behaviour: empirical study of the Taiwan stock market. *Applied Financial Economics* **23** 325-345.

[13] Cont, R., Kukanov, A. and Stoikov, S. (2014). The price impact of order book events. *Journal of Financial Econometrics* **12** 47-88.

[14] Corsi, F., Pirino, D. and Reno, R. (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics* **159** 276-288. MR2733121

[15] Czado, C. and Haug, S. (2010). An ACD-ECOGARCH(1,1) model. *Journal of Financial Econometrics* **8** 335-344.

[16] Drost, F. C. and Werker, B. J. M. (1996). Closing the GARCH gap: Continuous time GARCH modeling. *Journal of Econometrics* **74** 31-57.

MR1409034

[17] Eaves, J. and Williams, J. C. (2010). Are intraday volume and volatility U-shaped after accounting for public information? *American Journal of Agricultural Economics* **92** 212-227.

[18] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of inflation in the United Kingdom Inflation. *Econometrica* **50** 987-1007. MR0666121

[19] Engle, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica* **68** 1-22.

[20] Engle, R. F. and Sokalska, M. E. (2012). Forecasting intraday volatility in the US equity market. Multiplicative component GARCH. *Journal of Financial Econometrics* **10** 54-83.

[21] Fan, J. and Kim, D. (2019). Structured volatility matrix estimation for non-synchronized high-frequency financial data. *Journal of Econometrics* **209** 61-78. MR3913259

[22] Fan, J., Li, Y. and Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association* **107** 412-428. MR2949370

[23] Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association* **102** 1349-1362. MR2372538

[24] Gau, Y. and Hua, M. (2007). Intraday exchange rate volatility: ARCH, news and seasonality effects. *The Quarterly Review of Economics and Finance* **47** 135-158.

[25] Gerlach, R. H., Naimoli, A. and Storti, G. (2020). Time-varying parameters realized GARCH models for tracking attenuation bias in volatility dynamics. *Quantitative Finance* **20** 1849-1878. MR4159598

[26] Giot, P. (2005). Market risk models for intraday data. *European Journal of Finance* **11** 309-324.

[27] Glosten, L. R. and Harris, L. E. (1988). Estimating the components of the bid/ask spread. *Journal of Financial Economics* **21** 123-142.

[28] Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* **14** 71-100.

[29] Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711-732. MR1380810

[30] Hansen, P. R., Huang, Z. and Shek, H. (2012). Realized GARCH: A joint model of returns and realized measures of volatility. *Journal of Applied Econometrics* **27** 877-906. MR2993281

[31] Heston, S. L. (1993). A closed-form solution for options with stochastic volatilities with applications to bond and currency options. *The Review of Financial Studies* **6** 327-343. MR3929676

[32] Joseph, K. and Garcia, P. (2018). Intraday market effects in electronic soybean futures market during non-trading and trading hour announcements. *Applied Economics* **50** 1188-1202.

[33] Kalev, P. S., Liu, W., Pham, K. P. and Jarnecic, E. (2004). Public information arrival and volatility of intraday stock returns. *Journal of Banking & Finance* **28** 1441-1467.

[34] Kavajecz, K. (1999). A specialist's quoted depth and the limit order book. *The Journal of Finance* **54** 747-771.

[35] Keim, D. B. and Madhavan, A. (1996). The upstairs market for large-block transactions: Analysis and measurement of price effects. *Review of Financial Studies* **9** 1-36.

[36] Kim, D. and Fan, J. (2019). Factor GARCH-Itô models for high-frequency data with application to large volatility matrix prediction. *Journal of Econometrics* **208** 395-417. MR3913244

[37] Kim, D., Shin, M. and Wang, Y. (2022). Overnight GARCH-Itô volatility models. *Journal of Business & Economic Statistics*. MR4650456

[38] Kim, D. and Wang, Y. (2016). Unified discrete-time and continuous- time models and statistical inferences for merged low-frequency and high- frequency financial data. *Journal of Econometrics* **194** 220-230. MR3536977

[39] Klüppelberg, C., Lindner, A. and Maller, R. (2004). A continuous time GARCH process driven by a Levy process: Stationarity and second order behaviour. *Journal of Applied Probability* **41** 601-622. MR2074811

[40] Kraus, A. and Stoll, H. R. (1972). Price impacts of block trading on the New York Stock Exchange. *The Journal of Finance* **27** 569-588.

[41] Li, J., Todorov, V. and Tauchen, G. (2017). Robust jump regressions. *Journal of the American Statistical Association* **112** 332-341. MR3646575

[42] Li, J., Todorov, V. and Tauchen, G. (2017). Jump regressions. *Econometrica* **85** 173-195. MR3611769

[43] Li, Y., Mykland, P. A., Renault, E., Zhang, L. and Zheng, X. (2014). Realized volatility when sampling times are possibly endogenous. *Econometric Theory* **30** 580-605. MR3205607

[44] Li, Y., Xie, S. and Zheng, X. (2016). Efficient estimation of integrated volatility incorporating trading information. *Journal of Econometrics* **195** 33-50. MR3545292

[45] Mancini, C. (2004). Estimation of the characteristics of the jumps of a general Poisson-diffusion model. *Scandinavian Actuarial Journal* **1** 42-52. MR2045358

[46] Müller, G. (2010). MCMC Estimation of the COGARCH(1,1) Model. *Journal of Financial Econometrics* **8** 481-510.

[47] Mykland, P. A. and Zhang, L. (2012). The econometrics of high frequency data. *In Statistical Methods for Stochastic Differential Equations* 109-190. CRC Press. MR2976983

[48] Øksendal, B. (2003). Stochastic differential equations. Springer. MR2001996

[49] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components(with discussion). *Journal of the Royal Statistical Society: Series B* **59** 731-792. MR1483213

[50] Roll, R. (1984). Forecasting spot and forward prices in the international freight market. *The Journal of Finance* **39** 1127-1139.

[51] Schöbel, R. and Zhu, J. (1999). Stochastic volatility with an Ornstein-Uhlenbeck process: An extension. *European Finance Review* **3** 23-46.

[52] Scott, L. O. (1987). Option pricing when the variance changes randomly: Theory, estimation, and an application. *Journal of Financial and Quantitative analysis* **22** 419-438.

[53] Song, X., Kim, D., Yuan, H., Cui, X., Lu, Z., Zhou, Y. and Wang, Y. (2021). Volatility analysis with realized GARCH-Itô models. *Journal of Econometrics* **222** 393-410. MR4234824

[54] Taylor, S. J. (2007). *Modelling financial time series*, 2nd ed. World Scientific, Lancaster University.

[55] Wang, Y. and Zhou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *Annals of Statistics* **38** 943-978. MR2604708

[56] Wiggins, J. B. (1987). Option values under stochastic volatility: Theory and empirical estimates. *Journal of Financial Economics* **19** 351-372.