Statistica Sinica Preprint No: SS-2023-0020	
Title	Volatility Analysis with High-frequency and Low-frequency Historical Data, and Options-Implied Information
Manuscript ID	SS-2023-0020
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0020
Complete List of Authors	Huiling Yuan, Kexin Lu and Guodong Li
Corresponding Authors	Kexin Lu
E-mails	neithen@connect.hku.hk
Notice: Accepted version subject to English editing.	

Statistica Sinica

VOLATILITY ANALYSIS WITH HIGH-FREQUENCY AND LOW-FREQUENCY HISTORICAL DATA, AND OPTIONS-IMPLIED INFORMATION

Huiling Yuan[1], Kexin Lu*[2], and Guodong Li[2]

[1] *School of Statistics and Academy of Statistics and Interdisciplinary Sciences, and KLATASDS-MOE, East China Normal University*

[2] *Department of Statistics & Actuarial Science, The University of Hong Kong*

Abstract: In order to better forecast the volatility of underlying securities, this paper proposes a new model to integrate three important information sources, high-frequency and low-frequency historical observations, and options data, and the extracted options-implied information, rather than the option prices themselves, is used as an exogenous variable for simplicity. The quasi-maximum likelihood estimation is considered, and its asymptotic properties, as well as a hypothesis test, are further established. The proposed model is applied to the S&P 500 index as the underlying security, and the implied volatility index (VIX) is chosen as the options-implied information. Its superior performance can be observed in terms of out-of-sample forecast against many existing models.

Key words and phrases: Exogenous variable, Prediction performance, Quasi-

*Corresponding author.

maximum likelihood estimation, Simple volatility model.

1. Introduction

Volatility analysis has significant importance in modern finance and economics. A natural information source of the volatility is the historical data of a security, and they can be further classified into high- and low-frequency ones. Specifically, high-frequency data pertains to intra-day prices, such as tick-by-tick, 1-second and 5-minute data. Scholars commonly employ continuous-time Itô processes for high-frequency data, and the corresponding estimation can also be established. Examples include two-time scale realized volatility (TSRV) in Zhang et al. (2005), multi-scale realized volatility (MSRV) in Zhang (2006), kernel realized volatility (KRV) in Barndorff-Nielsen et al. (2009), pre-averaging realized volatility (PRV) in Jacod et al. (2009) and quasi-maximum likelihood estimation (QMLE) in Xiu (2010). On the other hand, low-frequency data refer to observed prices of a security over daily or longer time intervals, and many volatility models have been developed and evaluated for them; see, e.g., the autoregressive conditional heteroskedasticity (ARCH) in Engle (1982) and the generalized autoregressive conditional heteroskedasticity (GARCH) in Bollerslev (1986).

Intuitively, the relationship between high- and low-frequency historical data suggests the need of an integrated approach. In fact, many models have already been proposed for such integration, and they include the heteroge-

neous autoregressive model for realized volatility (HAR-RV) in Corsi (2009), the realized GARCH model (Hansen et al., 2012), the high-frequency-based volatility model (Shephard and Sheppard, 2010), the multiplicative error model (Engle and Gallo, 2006) and the GARCH-Itô model (Kim and Wang, 2016). The first four models incorporate daily realized volatility estimators directly into low-frequency econometric models as exogenous variables, while the GARCH-Itô model offers a more comprehensive approach to combine both high- and low-frequency data. Specifically, it integrates a GARCH model for low-frequency data at integer time points with a continuous-time Itô process between two consecutive points. As a result, the GARCH-Itô model is expected to be able to utilize the high-frequency data more effectively than the other four models.

In the meanwhile, option prices are another important information source for a security's volatility. Different from historical data which provides information from the past, it is forward-looking as it reflects expectation for the future. The most renowned measure derived from option prices is the implied volatility and, when being inputted in the Black-Scholes option pricing model, it yields a theoretical value equal to the current market price of the options. The implied volatility index (VIX) is a weighted average of implied volatilities for certain options on a specific index. It serves

as an indicator of investors' expectation regarding future economic conditions and market volatility. Numerous research efforts have been spent on constructing volatility models that combine low-frequency data and implied volatility information. For example, Blair et al. (2001) and Koopman et al. (2005) integrated the implied volatility as well as the realized volatility into low-frequency ARCH and GARCH models. Ni et al. (2008) and Chang et al. (2010) integrated the net demand for volatility of non-market makers and foreign institutional investors into linear volatility forecasting models.[1] More recently, Martin et al. (2021) demonstrated that the inclusion of the VIX in GARCH models can improve the volatility forecasts; see also Pan et al. (2019) and Pati et al. (2018).

On the other hand, many models are considered to combine all the three information sources: high-frequency, low-frequency and options data. Song et al. (2021) expanded the GARCH-Itô model to realized GARCH-Itô models with options-implied information, based on the linear relationship between the nonparametric volatility estimator (Todorov, 2019) and the conditional daily integrated volatility. Moreover, Yuan et al. (2022) found

[1]For each option, there is a difference between the number of bought contracts and sold contracts. The net demand for volatility is the vega-weighted sum of these differences across all options.

that there exists a dynamic equation between the options-implied information and security's conditional variance, leading to the GARCH-Itô-IV model, which is a "complete" specification of the joint dynamics of returns, latent instantaneous volatilities and options-implied volatility.

Motivated by the aforementioned researches, we treat the options-implied information as an exogenous variable, and hence a simpler volatility model, named GARCH-Itô-OI model, is proposed in Section 2. Its volatility properties are also discussed, and the options-implied information is assumed to have a positive and time-dependent influence on high-frequency instantaneous volatility in subsequent time intervals. Section 3 considers the quasi-maximum likelihood estimation, and its asymptotic properties, as well as a hypothesis test, are also established. Section 4 conducts simulation experiments to evaluate the finite-sample performance of the proposed methodology. Section 5 considers the S&P 500 index as the underlying security and utilizes the VIX as the implied information derived from options on the S&P 500 index. The proposed model is demonstrated to outperform existing volatility models in terms of forecasting accuracy and measured by mean squared error and quasi-likelihood. Section 6 gives a short conclusion, and all technical proofs are relegated to the Supplementary Material.

2. Unified GARCH-Itô-OI model

The GARCH-Itô model (Kim and Wang, 2016) unifies both high- and low-frequency historical data. Specifically, it embeds a standard GARCH(1,1) model into a continuous-time Itô process:

$$\begin{aligned} dX_t =& \mu_t dt + \sigma_t dB_t, \\ \sigma_t^2 =& \sigma_{[t]}^2 + (t-[t])\left\{\omega + (\gamma-1)\sigma_{[t]}^2\right\} + \beta\left(\int_{[t]}^{t} \sigma_s dB_s\right)^2, \end{aligned}$$

where X_t is the log price of an asset at time $t \in [0, +\infty)$, μ_t is a drift term, $[t]$ denotes the integer part of t except that $[t] = t-1$ when t is an integer, B_t is a standard Brownian motion with respect to a filtration $\{\mathcal{F}_t\}$, and σ_t^2 is a volatility process adapted to $\{\mathcal{F}_t\}$.

Due to its strong predictive power in forecasting the volatility of a security, the options-implied information is incorporated into the GARCH-Itô model, and this leads to the following GARCH-Itô-OI model.

Definition 1. A log price X_t with $t \in [0, +\infty)$ is said to follow a unified GARCH-Itô-OI model, if it satisfies that

$$\begin{aligned} dX_t =& \mu_t dt + \sigma_t dB_t, \\ \sigma_t^2 =& \sigma_{[t]}^2 + (t-[t])\left\{\omega + (\gamma-1)\sigma_{[t]}^2 + \alpha O_{[t]}\right\} + \beta\left(\int_{[t]}^{t} \sigma_s dB_s\right)^2, \end{aligned} \tag{2.1}$$

where $O_{[t]}$ is the $\{\mathcal{F}_{[t]}\}$-adapted exogenous options-implied information at integer time $[t]$, and $\theta = (\omega, \beta, \gamma, \alpha)$ is the parameter vector.

Remark 1. The proposed model involves the options-implied information at integer time points only, and we may alternatively consider the instantaneous version, i.e. $O_{[t]}$ is replaced by O_t. However, the options-implied information will be involved in a complicated form at the corresponding low-frequency GARCH model at the forthcoming Proposition 1, while this paper aims at a simple volatility model with the options-implied information being involved as an exogenous variable. To this end, we keep the proposed model at Definition 1 unchanged, while the model with instantaneous options-implied information is left for future research.

Remark 2. Based on practical considerations, we make certain assumptions on the options-implied information and its influence on the instantaneous volatility for the above model. Firstly, the options-implied information is assumed to be positive. Secondly, both the options-implied information and the past instantaneous volatility have time-dependent influences on the instantaneous volatility in the next time interval. This assumption acknowledges the fact that the impact of these variables may vary over time and hence can capture the dynamic of volatility forecasting. Thirdly, as the options-implied information is forward-looking, it is reasonable to assume

that it has more predictive power for the farther future, i.e. $\alpha > 0$. Lastly, the assumption that γ is less than one acknowledges the fact that, as time moves forward, the predictive power of past volatility decreases.

Consider the integrated volatility over consecutive integers, i.e. $\int_{i-1}^{i} \sigma_t^2 dt$ with $i \in \mathbb{N}$, where $\mathbb{N}$ is the set of all non-negative integers. We next state some of its properties.

Proposition 1. *(a) Under the GARCH-Itô-OI model, it holds that, for any* $i \in \mathbb{N}$ *and* $0 < \beta < 1$,

$$\int_{i-1}^{i} \sigma_t^2 dt = g_i(\theta) + D_i,$$

where

$$g_i(\theta) = \omega^g + \gamma g_{i-1}(\theta) + \beta^g Z_{i-1}^2 + \eta^g O_{i-1} + \xi^g O_{i-2} \tag{2.2}$$

with the parameters being given below

$$\begin{aligned} \omega^g &= \beta^{-1}(e^\beta - 1)\omega, \quad \beta^g = \beta^{-1}(\gamma - 1)(e^\beta - 1 - \beta) + e^\beta - 1, \\ \eta^g &= \beta^{-2}(e^\beta - 1 - \beta)\alpha, \quad \xi^g = \left[\beta^{-1}(e^\beta - 1) - \beta^{-2}(e^\beta - 1 - \beta)\right]\alpha, \\ \theta^g &= (\omega^g, \gamma, \beta^g, \eta^g, \xi^g), \end{aligned} \tag{2.3}$$

and

$$D_i = 2\int_{i-1}^{i} (e^{(i-t)\beta} - 1) \int_{i-1}^{t} \sigma_s dB_s \sigma_t dB_t,$$

is a martingale difference sequence.

(b) For any $i \in \mathbb{N}$ *and* $0 < \beta < 1$*, it holds that*

$$E\left[\int_{i-1}^{i} \sigma_t^2 dt \middle| \mathcal{F}_{i-1}\right] = g_i(\theta).$$

(c) For $0 < \beta^g + \gamma < 1$ *and* $i \in \mathbb{N}$*, it holds that*

$$E[g_i(\theta)] = \frac{\omega^g + (\eta^g + \xi^g)E[O_i]}{1 - \beta^g - \gamma} \quad \textit{and}$$

$$E[\sigma_i^2] = \frac{\omega(1 - \beta^g - \gamma) + \beta\omega^g + [\alpha(1 - \beta^g - \gamma) + \beta(\eta^g + \xi^g)]\, E[O_i]}{(1 - \beta^g - \gamma)(1 - \gamma)},$$

where ω^g*,* β^g*,* η^g *and* ξ^g *are defined in* (2.3)*.*

From (2.1), taking $\mu_t = 0$ for simplicity, the low-frequency log return $X_i - X_{i-1}$ will have a conditional variance equal to $E[Z_i^2|\mathcal{F}_{i-1}^{LF}]$, where $Z_i = \int_{i-1}^{i} \sigma_t dB_t$ and $\mathcal{F}_i^{LF} = \sigma(X_i, O_i, X_{i-1}, O_{i-1}, \cdots)$. Through the link at (2.2) between high- and low-frequency data, the above proposition implies that $X_i - X_{i-1}$ follows a GARCH(1,1) model with two exogenous variables,

$$E[Z_i^2|\mathcal{F}_{i-1}^{LF}] = \omega^g + \gamma E[Z_{i-1}^2|\mathcal{F}_{i-2}^{LF}] + \beta^g Z_{i-1}^2 + \eta^g O_{i-1} + \xi^g O_{i-2}. \qquad (2.4)$$

Model (2.4) embeds the low-frequency GARCH model with implied volatility (GARCH-IV) model (Koopman et al., 2005), which has already been shown by empirical studies to possess more accurate volatility forecasts compared to the standard GARCH model. More importantly, Definition 1 provides a high-frequency dynamic structure for (2.4) and builds a bridge between high-frequency and low-frequency parameters by (2.2) and (2.3).

3. Parameter estimation for GARCH-Itô-OI models

3.1 Quasi-maximum likelihood estimation

We first give the notations for the three information sources. Let X_t represent the underlying log price following the GARCH-Itô-OI model in Definition 1. The low-frequency data consist of observed true log prices at integer times, denoted by X_i with $0 \leq i \leq n$. The options-implied information is computed at integer times based on the true prices of options, and it is denoted by O_i with $0 \leq i \leq n$. Finally, the high-frequency data comprise log prices observed at $t_{i,j}$ with $0 \leq i \leq n$ and $j = 0, 1, \ldots, m_i + 1$, and these time points correspond to high-frequency intervals during the i-th period satisfying $i - 1 = t_{i,0} < t_{i,1} < \ldots < t_{i,m_i} < t_{i,m_i+1} = t_{i+1,0} = i$.

Different from low-frequency data and options-implied information, the observed high-frequency log prices are contaminated by micro-structure noise, making the true log prices unobservable. To address this issue, we assume that the observed high-frequency log prices $Y_{t_{i,j}}$ follow a simple additive noise model:

$$Y_{t_{i,j}} = X_{t_{i,j}} + \epsilon_{t_{i,j}}, \tag{3.5}$$

where $\epsilon_{t_{i,j}}$ is the micro-structure noise independent of $X_{t_{i,j}}$, and for each i, the noise sequence $\epsilon_{t_{i,j}}$'s with $j = 1, \cdots, m_i$ are independent and identically

distributed (*i.i.d.*) with mean zero and variance a_ϵ^2.

During the i-th period, the integrated volatility can be estimated by using high-frequency financial data. From low-frequency viewpoints, the estimated integrated volatilities can be treated as "observations" for instantaneous volatility. Similar to the quasi-likelihood function in Kim and Wang (2016), the quasi-likelihood function with options-implied information can be defined as follows,

$$\widehat{L}_{n,m}(\theta) = -\frac{1}{2n}\sum_{i=1}^{n}\log(g_i(\theta)) - \frac{1}{2n}\sum_{i=1}^{n}\frac{RV_i}{g_i(\theta)}, \tag{3.6}$$

where $g_i(\theta)$ is given at (2.2), and the realized volatility RV_i can be TSRV, MSRV, KRV, PRV or QMLE. From Proposition 1, the integrated volatility over the i-th period, $\int_{i-1}^{i}\sigma_t^2 dt$, equals to the sum of $g_i(\theta_0)$ and a martingale difference D_i, where $\theta_0 = (\omega_0, \beta_0, \gamma_0, \alpha_0)$ is the true parameter vector. Since the effect of martingale difference sequences are asymptotically negligible, the terms with D_i are dropped from the likelihood function $\widehat{L}_{n,m}(\theta)$. Moreover, the initial values of σ_0^2 and O_0 are needed to calculate $\widehat{L}_{n,m}(\theta)$. As a result, the quasi-maximum likelihood estimator can be defined below,

$$\widehat{\theta} = \arg\max_{\theta\in\Theta}\widehat{L}_{n,m}(\theta),$$

where Θ is the parameter space.

3.2 Asymptotic properties

This subsection establishes the consistency and asymptotic distribution for the proposed estimator $\widehat{\theta} = (\hat{\omega}, \hat{\beta}, \hat{\gamma}, \hat{\alpha})$. We first give some notations. For a matrix $A = (A_{i,j})_{i,j=1,\ldots,k}$ and a vector $a = (a_1, \ldots, a_k)$, their max norms are defined as $||A||_{\max} = \max_{i,j} |A_{i,j}|$ and $||a||_{\max} = \max_i |a_i|$, respectively. Given a random variable X, let its L_p norm be $||X||_{L_p} = \{E[|X|^p]\}^{1/p}$ for $p \geq 1$. Moreover, let C be a generic positive constant with its value being free of θ, n and m_i, and it may change from lines to lines.

Assumption 1. (a) The parameter space $\Theta = \{\theta = (\omega, \beta, \gamma, \alpha) : \omega_l < \omega < \omega_u, \beta_l < \beta < \beta_u, \gamma_l < \gamma < \gamma_u, \alpha_l < \alpha < \alpha_u, \gamma + \beta^g < 1\}$, where $\omega_l, \omega_u, \beta_l, \beta_u, \gamma_l, \gamma_u, \alpha_l$ and α_u are known positive constants, and β^g is given in (2.3).

(b) The options-implied information $\{O_i \geq 0 : i \in \mathbb{N}\}$ is uniformly bounded.

(c) One of the following two conditions is satisfied.

(c1) $E\left[Z_i^4 | \mathcal{F}_{i-1}\right] / g_i^2(\theta_0) \leq C$ a.s. for any $i \in \mathbb{N}$.

(c2) There exists a positive constant δ such that $E\left[(Z_i^2 / g_i(\theta_0))^{2+\delta}\right] \leq C$ for any $i \in \mathbb{N}$.

(d) $\{|D_i| : i \in \mathbb{N}\}$ is uniformly integrable.

(e) $\{D_i, Z_i^2\}$ is a stationary and ergodic process.

(f) Let $m = \sum_{i=1}^{n} m_i/n$. We have $C_1 m \leq m_i \leq C_2 m$, and $\sup_{1\leq j\leq m_i} |t_{i,j} - t_{i,j-1}| = O(m^{-1})$ and $n^2 m^{-1} \to 0$ as $m, n \to \infty$.

(g) $\sup_{i\in\mathbb{N}} \left\| RV_i - \int_{i-1}^{i} \sigma_t^2 dt \right\|_{L_{1+\delta}} \leq C \cdot m^{-1/4}$ for some $\delta > 0$.

(h) For any $i \in \mathbb{N}$, $E\left[RV_i | \mathcal{F}_{i-1}\right] \leq C \cdot E[\int_{i-1}^{i} \sigma_t^2 dt | \mathcal{F}_{i-1}] + C$ a.s.

In comparison to Assumption 1 in Kim and Wang (2016), we add an additional item (b) on the options-implied information, which is easily satisfied in general, while item (a) and (c)-(h) have been used in Kim and Wang (2016). Specifically, items (a)-(e) within Assumption 1 pertain to the low-frequency aspect of the model, while items (f)-(h) are associated with the high-frequency component.

Theorem 1. *(a) Under Assumption 1 (a), (b), (d) and (f)-(g), there exists a unique maximizer θ_0 of*

$$L_n(\theta) = -\frac{1}{2n}\sum_{i=1}^{n} \log(g_i(\theta)) - \frac{1}{2n}\sum_{i=1}^{n} \frac{g_i(\theta_0)}{g_i(\theta)},$$

and $\widehat{\theta} \to \theta_0$ in probability as $m, n \to \infty$.

(b) Under Assumption 1 (a)-(d) and (f)-(h), we have

$$\left\| \widehat{\theta} - \theta_0 \right\|_{\max} = O_p\left(m^{-1/4} + n^{-1/2}\right).$$

Theorem 1 shows that $\widehat{\theta}$ has the same convergence rate as the estimator in GARCH-Itô model (Kim and Wang, 2016). In other words, the options-implied information has no significant effect on the convergence rate.

Theorem 2. *Suppose that Assumption 1 holds. If* $m, n \to \infty$, *then*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, B^{-1}AB^{-1}),$$

where

$$A = E\left[\frac{\partial g_1(\theta)}{\partial \theta}\frac{\partial g_1(\theta)}{\partial \theta^T}\bigg|_{\theta=\theta_0} g_1^{-4}(\theta_0)\int_0^1 (e^{\beta_0(1-t)} - 1)^2 (X_t - X_0)^2 \sigma_t^2 dt\right],$$
$$B = \frac{1}{2}E\left[\frac{\partial g_1(\theta)}{\partial \theta}\frac{\partial g_1(\theta)}{\partial \theta^T}\bigg|_{\theta=\theta_0} g_1^{-2}(\theta_0)\right].$$

The asymptotic variance presented in Theorem 2 bears a resemblance to that of Kim and Wang (2016), with the inclusion of the options-implied information influencing through $g_1(\theta)$. Moreover, the proposed GARCH-Itô-OI model in Definition 1 will reduce to a GARCH-Itô model when $\alpha = 0$, and it hence is of interest to check the significance of $\hat{\alpha}$, i.e. the estimator of α. We may also construct a likelihood ratio test for it.

3.3 Hypothesis test

In financial practices, the primary focus of investors often lies on the low-frequency parameter vector $\theta^g = (\omega^g, \gamma, \beta^g, \eta^g, \xi^g)$ and their statistical inferences, although Theorem 2 establishes the asymptotic normality for the high-frequency parameter vector θ. Therefore, in this subsection, we discuss hypothesis test for θ^g.

From (2.2) and (2.3), the parameters $\omega^g, \gamma, \beta^g, \eta^g$ and ξ^g are functions of ω, β, γ and α. Let $\theta^g \equiv f(\theta) = (f_1(\theta), f_2(\theta), f_3(\theta), f_4(\theta), f_5(\theta))$, where $f_1(\theta) = \beta^{-1}\left(e^{\beta}-1\right)\omega$, $f_2(\theta) = \gamma$, $f_3(\theta) = \beta^{-1}(\gamma-1)\left(e^{\beta}-1-\beta\right)+e^{\beta}-1$, $f_4(\theta) = \beta^{-2}\left(e^{\beta}-1-\beta\right)\alpha$ and $f_5(\theta) = \left[\beta^{-1}\left(e^{\beta}-1\right)-\beta^{-2}\left(e^{\beta}-1-\beta\right)\right]\alpha$. Denote $f(\widehat{\theta})$ by $\widehat{\theta}^g$.

Building on the asymptotic distribution of $\widehat{\theta}$ in Theorem 2, and with the aid of Slutsky's theorem and the delta method, when $\partial f(\theta)/\partial\theta|_{\theta=\theta_0} \neq 0$, we can obtain that, for any constant vector $c \in \mathbb{R}^5$,

$$Z_f = \frac{\sqrt{n}(c^{\top}\widehat{\theta}^g - c^{\top}\theta_0^g)}{\sqrt{c^{\top}\nabla f(\widehat{\theta})^T(\widehat{B}^{-1}\widehat{A}\widehat{B}^{-1})^{-1}\nabla f(\widehat{\theta})c}} \xrightarrow{d} \mathcal{N}(0,1), \tag{3.7}$$

where $\nabla f(\widehat{\theta}) = \partial f(\theta)/\partial\theta|_{\theta=\widehat{\theta}}$. Matrices A and B can be consistently estimated by

$$\widehat{A} = \frac{1}{4n^2}\sum_{i=1}^{n}\frac{\partial\widehat{\ell}_i(\theta)}{\partial\theta}\frac{\partial\widehat{\ell}_i(\theta)}{\partial\theta^T} \quad \text{and} \quad \widehat{B} = -\frac{\partial^2\widehat{L}_{n,m}(\theta)}{\partial\theta\partial\theta^T},$$

respectively, where $\widehat{\ell}_i(\theta)$ and $\widehat{L}_{n,m}(\theta)$ are given in the Supplementary Material. Then, using the Z-statistics Z_f in (3.7), the hypothesis test for linear combinations of $\widehat{\theta}^g$ can be conducted based on the standard normal distribution. For example, a hypothesis test for $\mathrm{H}_0 : \omega^g = 0$ can be conducted by setting $c = (1,0,0,0,0)^{\top}$.

4. Simulation studies

4.1 Finite-sample performance of the proposed estimator

To assess the finite-sample performance of the proposed estimators, we simulate the log prices $X_{t_{i,j}}$ and instantaneous volatilities $\sigma^2_{t_{i,j}}$ from the GARCH-Itô-OI model, where $t_{i,j} = i - 1 + j/m$ with $1 \leq i \leq n$ and $1 \leq j \leq m$. The number of high-frequency observations, m, in a trading day is assumed to be the same throughout the whole period for simplicity (Shin et. al., 2021). The true parameter vector is set to be $\theta_0 = (0.2, 0.3, 0.4, 0.1)^\top$. The incremental positions ΔB_t of the Brownian motion B_t are generated from normal distributions with mean zero and variance $1/m$. Three cases of distributions are considered for the options-implied information O_i: (a) $O_i \sim \mathcal{N}(0.6, 0.01)$, (b) $O_i \sim \mathcal{U}(0, 0.5)$, and (c) $O_i \sim$log-normal $(0.6, 0.01)$. We further set the initial log price $X_0 = 10$, and the initial options-implied information $O_0 = 0.5$ for case (a), $O_0 = 0.25$ for case (b) and $O_0 = 1.2$ for case (c), respectively. According to Proposition 1 (c), we can compute the instantaneous volatilities $\sigma_0^2 = 0.87$ for case (a), $\sigma_0^2 = 0.75$ for case (b) and $\sigma_0^2 = 1.67$ for case (c). The observed high-frequency data $Y_{t_{i,j}}$ are obtained through model (3.5) by simulating the micro-structure noise $\epsilon_{t_{i,j}}$ from $\mathcal{N}(0, 0.001^2)$. The sample size n varies in $\{125, 250, 500\}$, and the number

of observations m varies in $\{78, 390, 23400\}$. We generate 1000 replications for each combination of n and m. The multi-scale realized volatility estimator (MSRV) in Zhang (2006) is used to estimate the integrated volatility. Specifically, for the i-th period,

$$\text{MSRV}_i = \sum_{k=1}^{[\sqrt{m}]} a_k \cdot \text{MSRV}_i^{K_k} + \zeta(\text{MSRV}_i^{K_1} - \text{MSRV}_i^{K[\sqrt{m}]}),$$

where $\text{MSRV}_i^K = K^{-1} \sum_{r=K}^{m} \left[\text{Y}(t_{i,r}) - \text{Y}(t_{i,r-K})\right]^2, K_k = [\sqrt{m}] + k,\ a_k = \frac{12(k+[\sqrt{m}])(k-[\sqrt{m}]/2-1/2)}{[\sqrt{m}](m-1)}$, and $\zeta = \frac{2[\sqrt{m}]([\sqrt{m}]+1)}{(m+1)([\sqrt{m}]-1)}$.

The mean square errors (MSE) of the parameter estimator $\widehat{\theta}$ for different combinations of n and m are presented in Table 1. It is worth noting that, as both n and m increase, the MSEs consistently decrease across all three distributions of the options-implied information O_i. Moreover, for the log-normal distribution, the MSEs are relatively larger compared to those with the normal and uniform distributions, showing the influence of heavy-tailedness on the estimator. These findings collectively demonstrate the favorable finite-sample performance of the proposed estimator and provide further validation for its asymptotic behaviors.

Figure 1 gives the Q-Q plots of estimated parameters with $n = 500$ and $m = 23400$ for three cases of distributions. It can be seen that the estimated parameters closely align with the normal theoretical quantiles,

Table 1: Mean square errors (MSE) of $\widehat{\theta}$ based on 1000 replications. The options-implied information follows $\mathrm{O}_i \sim \mathcal{N}(0.6, 0.01)$, $\mathcal{U}(0, 0.5)$ or log-normal$(0.6, 0.01)$.

MSE$\times 10^{-3}$													
n	m	$\mathrm{O}_i \sim \mathcal{N}(0.6, 0.01)$				$\mathrm{O}_i \sim \mathcal{U}(0, 0.5)$				$\mathrm{O}_i \sim$ log-normal$(0.6, 0.01)$			
		ω	β	γ	α	ω	β	γ	α	ω	β	γ	α
125	78	40.47	6.08	17.73	100.61	5.77	6.29	17.70	27.26	102.37	6.00	15.79	28.68
	390	22.54	2.69	7.52	58.97	2.39	2.76	7.05	16.62	55.98	2.80	7.70	16.13
	23400	9.44	0.90	2.20	24.59	0.87	0.89	2.39	6.00	19.58	0.89	2.44	5.74
250	78	18.28	3.66	6.79	45.62	3.31	3.89	7.29	15.21	59.82	4.10	7.40	16.04
	390	12.22	1.48	3.23	30.87	1.27	1.47	3.32	8.65	34.66	1.51	3.41	10.15
	23400	3.89	0.45	1.10	10.27	0.44	0.51	1.15	3.65	13.83	0.46	1.07	4.05
500	78	9.73	2.91	3.57	23.22	2.21	2.96	3.42	7.61	39.37	3.12	3.53	11.02
	390	5.32	0.87	1.60	13.54	0.71	0.90	1.59	4.89	23.11	0.90	1.59	6.72
	23400	2.06	0.25	0.56	5.31	0.23	0.23	0.57	1.84	8.38	0.26	0.61	2.50

indicating the desirable asymptotic normality.

4.2 Testing performance

To check the asymptotic normality of $\widehat{\theta}^g = (\widehat{\omega}^g, \widehat{\gamma}, \widehat{\beta}^g, \widehat{\eta}^g, \widehat{\xi}^g)$, we calculate the Z-statistics defined in Section 3.3. The data generating process is the same as that in the previous subsection with $m = 23400$ and $n = 500$, and the Q-Q plots of each estimated parameter are presented in Figure 2. It can

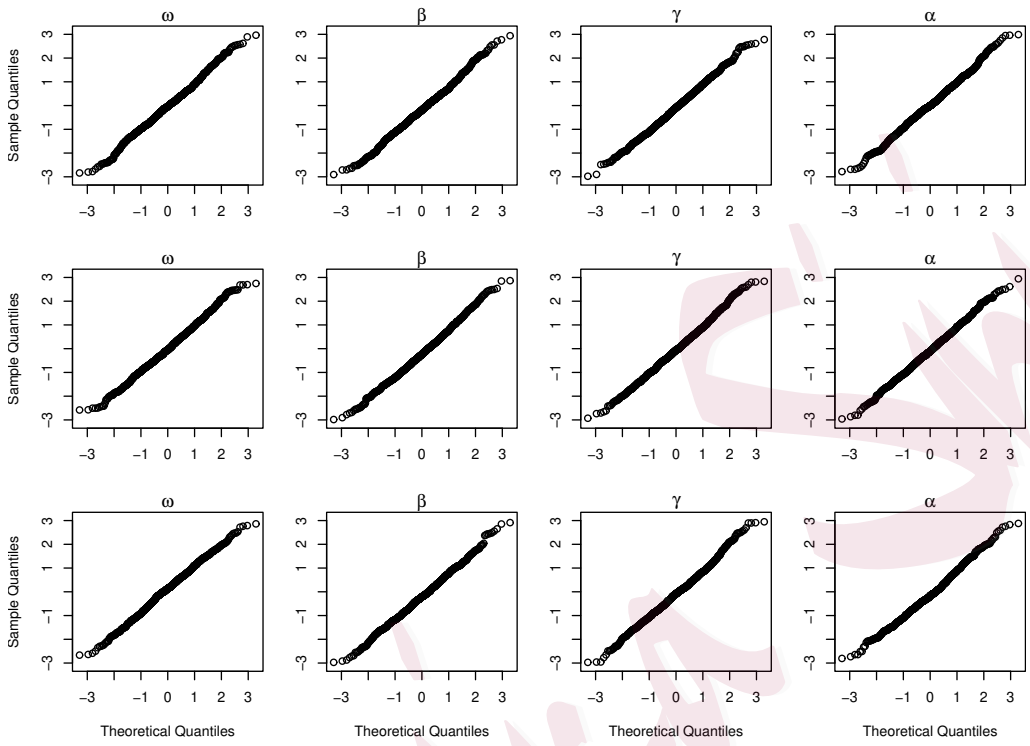

Figure 1: Q-Q plots for $\widehat{\theta} = (\widehat{\omega}, \widehat{\beta}, \widehat{\gamma}, \widehat{\alpha})$ with $(n, m) = (500, 23400)$. The options-implied information $\mathrm{O}_i \sim \mathcal{N}(0.6, 0.01)$ in the upper panel, $\mathcal{U}(0, 0.5)$ in the middle panel or log-normal$(0.6, 0.01)$ in the lower panel.

be seen that all sample quantiles are close to the corresponding theoretical ones, consistent with the theoretical conclusions in Section 3.3.

4.3 Prediction performance

This subsection evaluates the prediction performance of GARCH-Itô-OI models under model misspecification. Specifically, the log prices obey the

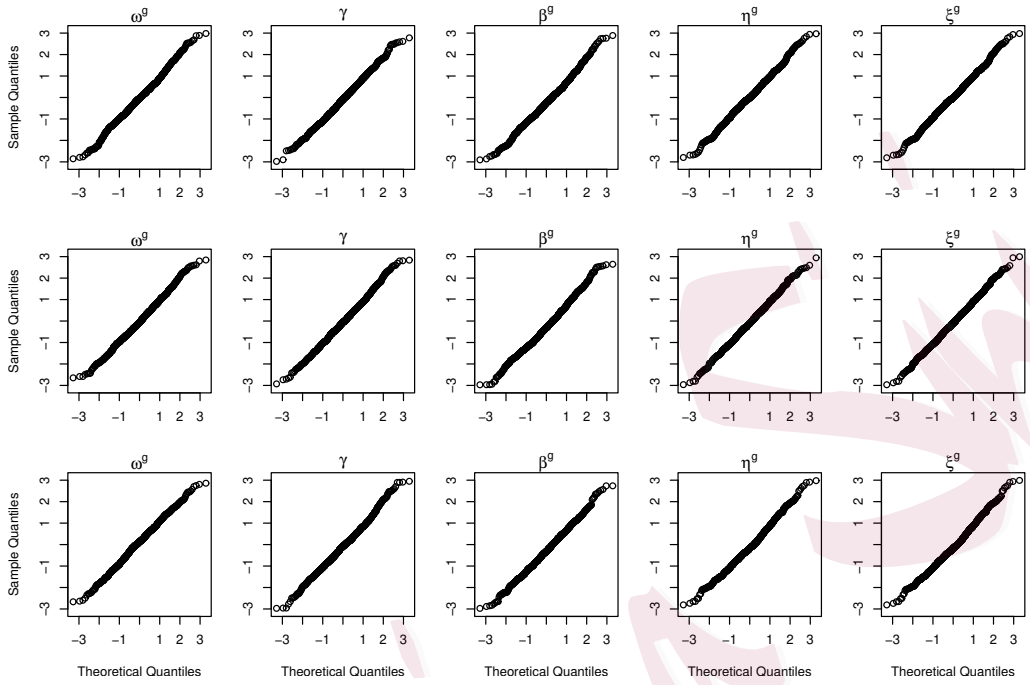

Figure 2: Q-Q plots for $\widehat{\theta}^g = (\widehat{\omega}^g, \widehat{\gamma}, \widehat{\beta}^g, \widehat{\eta}^g, \widehat{\xi}^g)$ with $(n, m) = (500, 23400)$. The options-implied information $\mathrm{O}_i \sim \mathcal{N}(0.6, 0.01)$ in the upper panel, $\mathcal{U}(0, 0.5)$ in the middle panel or log-normal$(0.6, 0.01)$ in the lower panel.

following Heston model,

$$\begin{aligned} dS_t &= rS_t dt + \sqrt{V_t} S_t dW_t, \\ dV_t &= (a - bV_t)dt + \gamma\sqrt{V_t} dW_t^*, \end{aligned} \tag{4.8}$$

where W_t and W_t^* are two standard Brownian motions with correlation coefficient ρ, and the volatility V_t acts as the options-implied volatility. We set $r = 0$, $a = 0.01$, $b = 0.001$, $\gamma = 0.075$, $\rho = -0.2$, $S_0 = 50$, $V_0 = 0.05$, and the number of high-frequency data $m = 78$. We generate 1000 replications,

and there are 101 days in each replication. The first 100 days are used for estimating parameters, and the 101st day is for evaluating the out-of-sample forecast performance.

Table 2: The AMAPE ($\times 10^{-1}$) , MAE ($\times 10^{-4}$), MSE ($\times 10^{-8}$) and QLIKE ($\times 10^{-1}$) of eight models. Lowest values are highlighted in boldface.

Model	AMAPE	MAE	MSE	QLIKE
GARCH-Itô-OI	**1.84**	**1.71**	**5.76**	**1.12**
GARCH-Itô-IV	1.91	1.77	6.07	1.25
GARCH-Itô	1.93	1.79	6.14	1.27
HAR-RV	1.96	1.84	6.20	1.28
RGARCH-Itô-NV	2.01	1.94	6.38	1.29
RGARCH-Itô	2.08	2.02	6.46	1.31
RGARCH-RV	2.23	2.18	8.05	1.58
GARCH-IV	2.31	2.25	8.62	2.38

As a comparison, we also fit seven existing volatility models in the literature: the GARCH-IV, realized GARCH-RV (RGARCH-RV), HAR-RV, GARCH-Itô, GARCH-Itô-IV, Realized GARCH-Itô (RGARCH-Itô), and Realized GARCH-Itô with estimated options (RGARCH-Itô-NV). Four

evaluation metrics are considered: the adjusted mean absolute percentage error (AMAPE), mean absolute error (MAE), mean squared error (MSE), and quasi-likelihood (QLIKE),

$$\text{AMAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\text{F}_i - \text{RV}_i|}{|\text{F}_i + \text{RV}_i|}, \quad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\text{RV}_i - \text{F}_i|,$$
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\text{RV}_i - \text{F}_i)^2, \quad \text{QLIKE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{RV}_i}{\text{F}_i} - \log\left(\frac{\text{RV}_i}{\text{F}_i}\right) - 1\right),$$

where N is the number of out-of-sample forecasts, RV_i is the true daily volatility, and F_i is the predicted volatility. The prediction results are listed in Table 2, and it can be seen that the proposed GARCH-Itô-OI model has the best performance among all competing models.

5. Empirical studies

This section applies the proposed GARCH-Itô-OI model to forecast the volatility of S&P 500. Specifically, we utilize five-minute returns, since five-minute is a commonly employed time interval in the literature (Liu et al., 2015), and this corresponds to $m = 78$ intraday observations in total. Overnight returns are excluded as they are often subject to jumps influenced by external factors. The corresponding daily closing prices are the low-frequency data, and the options-implied information is set to the VIX, which is quoted at the market closing on each trading day, while

S&P 500 index is the underlying asset. The MSRV is used to estimate the integrated volatility.

Our study encompasses three distinct time periods to comprehensively evaluate the performance of GARCH-Itô-OI models. The first period spans from January 2, 2015 to December 31, 2019, and it is exactly prior to the COVID-19 outbreak. This period is further divided into two segments: an in-sample period running from January 2, 2015 to December 31, 2018, and an out-of-sample period covering January 2, 2019 to December 31, 2019. The second period is from January 2, 2020 to August 21, 2023, and it is after the COVID-19 outbreak. The in-sample period is from January 2, 2020 to December 31, 2022, and the out-of-sample period is from January 2, 2023 to August 21, 2023. The third period encompasses the entire study duration, including both pre- and post-COVID-19 periods, i.e. it spans from January 2, 2015 to August 21, 2023. Within this period, the in-sample segment covers January 2, 2015 to December 31, 2021, while the out-of-sample segment is from January 2, 2022 to August 21, 2023.

5.1 Model estimation

We first conduct the quasi-maximum likelihood estimation in Section 3.1 to the in-sample segment from each of three periods, and the resulting

estimator can be denoted by $\widehat{\theta} = (\widehat{\omega}, \widehat{\beta}, \widehat{\gamma}, \widehat{\alpha})$. Note that, when $\alpha = 0$, the GARCH-Itô-OI model will reduce to a GARCH-Itô model, i.e. the options-implied information has no impact on forecasting the volatility. It hence is of interest to check the significance of α, and Theorem 2 can be directly used for it. As suggested by a reviewer, we alternatively consider a formal likelihood-ratio test (LRT) statistic,

$$\lambda_{\mathrm{LR}} = -2n \left(\widehat{L}_{n,m}(\widehat{\theta}_{\mathrm{GI}}) - \widehat{L}_{n,m}(\widehat{\theta}) \right),$$

where $\widehat{\theta}_{\mathrm{GI}}$ is a quasi-maximum likelihood estimator of the corresponding GARCH-Itô model, $\widehat{L}_{n,m}(\widehat{\theta})$ is denoted by (3.6), and it follows a χ_1 distribution as $n, m \to \infty$ under the null hypothesis of $\alpha = 0$. Table 3 gives the values of LRT statistics.

It can be seen that all LRT statistics have values greater than $\chi^2_{1,0.05} = 3.841$, indicating that the options-implied information is significant at the significance level of 0.05. In other words, the options-implied information plays an important role in forecasting the volatility of S&P 500.

We next evaluate the significance of low-frequency parameters $\theta^g = (\omega^g, \gamma, \beta^g, \eta^g, \xi^g)$ in terms of predicted volatility, and a rolling-window procedure is employed. Specifically, the ending point of historical data iterates in the out-of-sample period with the window size fixed at the number of in-sample period, and then one-step ahead prediction, denoted by y_i^*, is

Table 3: Likelihood-ratio test statistics for the presence of options-implied information at three time periods.

LRT statistics	first period	second period	whole period
λ_{LR}	151.32	23.46	181.40

conducted for each iteration. With the help of (2.2), we then perform the regression, $y_i^* = \omega^g + \gamma y_{i-1}^* + \beta^g Z_{i-1}^2 + \eta^g O_{i-1} + \xi^g O_{i-2} + \varepsilon_i$, and Table 4 presents the estimated parameters and their p-values.

It can be seen that almost all estimated parameters are significant. Especially, the estimators of η^g and ξ^g have much smaller p-values, and this further validate the necessity of involving options-implied information.

Table 4: Parameter estimation and p-values of the GARCH-Itô-OI model for the three time periods.

parameter	first period		second period		whole period	
	estimation	p-value	estimation	p-value	estimation	p-value
ω^g	3.02e-07	0.06	3.52e-07	0.05	8.39e-07	5.56e-08
γ	4.88e-01	$<$ 2e-16	4.92e-01	$<$ 2e-16	4.39e-01	$<$ 2e-16
β^g	8.02e-02	$<$2e-16	1.00e-01	$<$ 2e-16	8.67e-02	$<$ 2e-16
η^g	5.99e-02	$<$ 2e-16	3.26e-02	$<$ 2e-16	4.18e-02	$<$ 2e-16
ξ^g	2.77e-02	1.39e-15	3.85e-02	$<$ 2e-16	4.37e-02	$<$ 2e-16

5.2 Prediction performance

This subsection compares the forecasting performance of the GARCH-Itô-OI model with seven competing models: GARCH-Itô-IV, GARCH-Itô, HAR-RV, RGARCH-Itô-NV, RGARCH-Itô, RGARCH-RV, and GARCH-IV models in the three periods. The same rolling-window procedure in the previous subsection is used again. We utilize two robust loss functions, MSE and QLIKE, to assess the prediction performance of these models, as they provide a consistent ranking of volatility models with conditionally unbiased volatility proxy for prediction performance in real data (Patton, 2011a,b). All results are summarized in Table 5.

From Table 5, it is evident that the GARCH-Itô-OI model exhibits the best prediction performance among all competing models. This highlights the effectiveness of incorporating options information within the GARCH-Itô-OI model. In comparison to the GARCH-Itô-IV model, the simple structure of GARCH-Itô-OI models enables efficient utilization of the options information, leading to improved forecasting accuracy. By comparing the GARCH-Itô-OI and GARCH-Itô-IV against GARCH-Itô models, and RGARCH-Itô-NV against RGARCH-Itô models, we may conclude that models incorporating options information outperform their respective base models without options information. This demonstrates the advantage of

Table 5: MSEs ($\times 10^{-9}$) and QLIKEs ($\times 10^{-1}$) of the eight models during the three time periods. Lowest values are highlighted in boldface.

Model	first period		second period		whole period	
	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
GARCH-Itô-OI	**0.94**	**5.19**	**1.38**	**3.14**	**5.82**	**3.22**
GARCH-Itô-IV	1.28	5.87	1.55	3.32	6.07	3.45
GARCH-Itô	1.34	5.91	1.48	3.36	6.16	3.55
HAR-RV	1.35	5.95	1.48	3.28	5.88	3.34
RGARCH-Itô-NV	1.29	5.89	1.53	3.29	6.09	3.43
RGARCH-Itô	1.35	5.96	1.59	3.37	6.35	3.56
RGARCH-RV	2.62	8.89	1.73	4.27	7.95	5.09
GARCH-IV	4.01	11.23	8.59	6.74	8.33	5.28

including options information in volatility modeling. It is worth noting that the GARCH-IV model exhibits the worst forecasting performance, as it fails to make use of high-frequency data effectively. Interestingly, despite its simple linear structure, the HAR-RV model demonstrates superior prediction performance compared to most models. However, it still cannot beat the proposed GARCH-Itô-OI model.

Although the advantages of our model in Table 5, we may wonder whether such outperfromance is statistically significant. To this end, we first conduct the Diebold-Mariano (DM) test (Diebold, 1995) for the comparison. Specifically, the squared prediction error is defined as $e_i^2 = (y_i - \widehat{y}_i)^2$ with $1 \leq i \leq N$, where y_i and $\widehat{y}_i$ are the realized volatility and its predicted value, respectively. The DM test is to check whether the squared prediction errors from one model is smaller than those from the baseline model. Table 6 lists the p-values of DM tests with the null hypothesis that the model has a better forecasting performance than the proposed GARCH-Itô-OI model. It can be seen that all tests are rejected, i.e. the prediction power of our model is significantly higher than the seven competing models.

We also conduct the Model Confidence Set (MCS) test (Hansen et al., 2011) to compare the eight volatility models, and the forecasting performance is evaluated by $\ell_i = (y_i^2 - \widehat{y}_i^2)^2$. The full set contains the eight competing model, and the MCS procedure will eliminate them one by one. We also calculate the probability of each set of superior models, and the results are presented in Table 7. It can be seen that our model has a comparable performance with GARCH-Itô-IV and RGARCH-Itô-NV model for the first period of data, with HAR-RV models for the second period and with RGARCH-Itô-NV models for the whole period. However, for all three

Table 6: p-values of the DM test with the null hypothesis that the model has a smaller squared prediction errors than those from the proposed GARCH-Itô-OI model.

Models	first period	second period	whole period
GARCH-Itô-IV	4.73e-03	4.36e-02	2.49e-02
GARCH-Itô	3.39e-03	1.95e-02	2.48e-02
HAR-RV	2.09e-03	4.97e-02	4.49e-02
RGARCH-Itô-NV	4.42e-03	4.49e-02	4.54e-02
RGARCH-Itô	1.48e-03	1.05e-02	2.36e-02
RGARCH-RV	2.73e-10	6.10e-05	1.33e-08
GARCH-IV	<2.2e-16	< 2.2e-16	2.25e-10

periods of data, the proposed GARCH-Itô-OI model is always the best one.

Finally, we may wonder whether a volatility model can interpret all autocorrelation structure in the data. To this end, we first conduct a regression of the actual volatility on the predicted volatility, and then the autocorrelation function (ACF) plot can be used to check the autocorrelation dynamics in the residuals. The ACFs plots of residuals from each model for three time periods are presented in Supplementary Material, and

Table 7: Model ranks and probabilities of model confidence sets, i.e. P_{MCS}. The full set includes all eight models, and they are eliminated one by one according to model ranks. The GARCH-IV model can be eliminated with probability one.

Model	first period		second period		whole period	
	rank	P_{MCS}	rank	P_{MCS}	rank	P_{MCS}
GARCH-Itô-OI	1	1	1	1	1	1
GARCH-Itô-IV	2	1	4	0.84	4	0.60
GARCH-Itô	4	0.96	5	0.79	5	0.34
HAR-RV	5	0.88	2	1	3	0.99
RGARCH-Itô-NV	3	1	3	0.99	2	1
RGARCH-Itô	6	0.78	6	0.37	6	0.14
RGARCH-RV	7	0.62	7	0.18	7	0.13
GARCH-IV	eliminated	–	eliminated	–	eliminated	–

it can be seen that all the eight models can well capture autocorrelation structures within the S&P 500 index future's volatility.

6. Conclusion

The GARCH-Itô-OI model proposed in this paper integrates three information sources: high-frequency, low-frequency, and options data. In par-

ticular, the options-implied information is treated as an exogenous variable and assumed to have a positive and time-dependent influence on the high-frequency instantaneous volatility in next day with a clear and simple format. The quasi-maximum likelihood estimation for parameters are derived, and its asymptotic properties, as well as a hypothesis test, are also provided. Simulation experiments are conducted to analyze the finite-sample performance of the GARCH-Itô-OI model. It is demonstrated by the real analysis that options-implied information is an important source for volatility forecasting, and the proposed GARCH-Itô-OI model has better out-of-sample forecasting performances than all existing volatility models.

The proposed GARCH-Itô-OI model can be further extended in the following two directions. First, jump components can be integrated into the model to generate better volatility forecasts. Secondly, as different financial market information such as asymmetry and overnight volatility have been considered in the GARCH-Itô model (Kim et al. , 2022; Yuan et al. , 2022), it is of interest to combine them with the options-implied imformation into a more comprehensive GARCH-Itô model.

Supplementary Materials

The online Supplementary Material contains the proofs of the proposition and theorems, and autocorrelation function plots in empirical study.

Acknowledgements

We are deeply grateful to the editor, an associate editor and two anonymous referees for their valuable comments that led to the substantial improvement of the manuscript. Yuan's research was partially supported by Shanghai Pujiang Program (No. 23PJ1402400), China Postdoctoral Science Foundation (No. 2023M741190) and State Key Program of National Natural Science Foundation of China (No. 71931004). Li's research was partially supported by GRF grant 17313722 from the Hong Kong Research Grant Council and a Key Program grant 72033002 from National Natural Science Foundation of China.

References

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2009). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* **76**, 1481-1536.

Blair, B. J., Poon, S.H. and Taylor, S. J. (2001). Forecasting S&P 100 volatility : the incremental

information content of implied volatilities and high-frequency index returns. *Journal of Econometrics* **105**, 5-26.

Bollerslev T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307-327.

Chang, C. C., Hsieh, P. F. and Wang, Y. H. (2010). Information content of options trading volume for future volatility: Evidence from the Taiwan options market. *Journal of Banking & Finance* **34**, 174-183.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* **7**, 174-196.

Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**, 253-263.

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

Engle, R.F., Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* **131**, 3-27.

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

Hansen P. R., Lunde A. and Nason J. M. (2011). The model confidence set. *Econometrica* **79**, 453-497.

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

Kim D., 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.

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

Koopman, S.J., Jungbacker, B. and Hol, E. (2005). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance* **12**, 445-475.

Liu, L. Y., Patton, A. J. and Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics* **187**, 293-311.

Martin, V. L., Tang, C. and Yao, W. (2021). Forecasting the volatility of asset returns: The informational gains from option prices. *International Journal of Forecasting* **37**, 862-880.

Ni, S. X., Pan, J. and Poteshman, A. M. (2008). Volatility information trading in the option market. *The Journal of Finance* **63**, 1059-1091.

Pan, Z., Wang, Y., Liu, L. and Wang, Q. (2019). Improving volatility prediction and option valuation using VIX information: A volatility spillover GARCH model. *Journal of Futures*

Market **39**, 744-776.

Pati, P.C., Barai, P. and Rajib, P. (2018). Forecasting stock market volatility and information content of implied volatility index. *Applied Economics* **50**, 2552-2568.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* **160**, 246-256.

Patton, A. J. (2011). Data-based ranking of realised volatility estimators. *Journal of Econometrics* **161**, 284-303.

Shephard, N., Sheppard, K. (2010). Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* **25**, 197-231.

Shin, M., Kim, D., Wang, Y. and Fan, J. (2021). Factor and Idiosyncratic VAR-Itô volatility models for heavy-tailed high-frequency financial data. *arXiv preprint arXiv:2109.05227*.

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.

Todorov, V. (2019). Nonparametric spot volatility from options. *The Annals of Applied Probability*, 29, 3590-3636.

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics* **159**, 235-250.

Yuan, H., Sun, Y., Xu, L., Zhou, Y. and Cui, X. (2022). A New Volatility Model: GQARCH-Itô Model. *Journal of Time Series Analysis* **43**, 345-370.

Yuan, H., Zhou, Y., Zhang, Z. Y. and Cui, X. (2022). Volatility Analysis for GARCH-Itô Model with Option Data. *Canadian Journal of Statistics*, DOI: 10.1002/cjs.11746.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli* **12**, 1019-1043.

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

School of Statistics and Academy of Statistics and Interdisciplinary Sciences, and KLATASDS-MOE, East China Normal University.

E-mail: hlyuan@sfs.ecnu.edu.cn

Department of Statistics & Actuarial Science, The University of Hong Kong.

E-mail: neithen@connect.hku.hk

Department of Statistics & Actuarial Science, The University of Hong Kong.

E-mail: gdli@hku.hk