When Moving-Average Models Meet High-Frequency Data: Uniform Inference on Volatility∗

Rui Da†
University of Chicago

Dacheng Xiu‡
University of Chicago

This version: October 15, 2020

Abstract

We conduct inference on volatility with noisy high-frequency data. We assume the observed transaction price follows a continuous-time Itô-semimartingale, contaminated by a discrete-time moving-average noise process associated with the arrival of trades. We estimate volatility, as defined by the quadratic variation of the semimartingale, by maximizing the likelihood of a misspecified moving-average model, with its order selected based on the information criteria. Our inference is uniformly valid over a large class of noise processes whose magnitude and dependence structure vary with sample size. We show that the convergence rate of our estimator dominates \( n^{1/4} \) as noise vanishes, and is determined by the selected order of noise dependence when noise is sufficiently small. Our implementation guarantees positive estimates in finite samples. Finally, we provide consistent estimators of noise autocovariances as byproducts, which play a critical role in achieving uniformity.

Keywords: QMLE, Serially Correlated Noise, Small Noise, Model Selection, Uniformity

JEL Codes: C13, C14, C55, C58.

∗We are very grateful to the editor Ulrich Müller and three anonymous referees for valuable comments. We also benefit tremendously from discussions with Yacine Aït-Sahalia, Torben Andersen, Roland Bénabou, Tim Bollerslev, Bruce Hansen, Christian Hansen, Peter Hansen, Per Mykland, Andrew Patton, Zhongjun Qu, Jeffrey Russell, Neil Shephard, Ruey Tsay, and Ken West, as well as seminar and conference participants at the University of Wisconsin, Imperial College London, UCL, Indiana University, Aarhus University, Tilburg University, the 9th SoFiE annual conference, the CEME Young Econometricians Workshop at Duke University, Conference on High Frequency Financial Econometrics at Rutgers, and Market Microstructure and High Frequency Trading Conference at the University of Chicago. Xiu gratefully acknowledges the computing support from the Research Computing Center at the University of Chicago and financial support from the Fama-Miller Center and the IBM Faculty Scholar Fund at the University of Chicago Booth School of Business, as well as an Azure Research Grant from Microsoft.

†Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: rui.da@chicagobooth.edu.

‡Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: dacheng.xiu@chicagobooth.edu.
1 Introduction

In this paper, we develop a simple estimator of volatility using high-frequency data in the presence of serially correlated, heteroscedastic, and endogenous microstructure noise. More importantly, we propose uniformly valid inference over a large class of noise processes that allows for infinite-order autocorrelation and arbitrarily shrinking magnitude simultaneously.

Hansen and Lunde (2006) provide empirical evidence that microstructure noise is quite small in Dow Jones Industrial Average stocks. To improve efficiency, one can consider a test of whether noise is present (or rely on an informal volatility signature plot), then decide whether to use a noise-robust estimator or the more efficient realized volatility estimator (Andersen, Bollerslev, Diebold, and Labys (2003)), which assumes noise is absent. Standard (pointwise) inference for this pre-testing approach, however, provides a misleading picture of the actual finite-sample behavior. Moreover, assuming the noise exists, a follow-up issue is to determine its dependence structure. An estimator robust to noise with long-range temporal dependence could be inefficient if the actual noise is simply i.i.d.. To strike a more desirable tradeoff between efficiency and robustness, one can consider modeling the noise as a moving average process and adopting information criteria to determine its order of dependence. Nevertheless, model-selection mistakes are inevitable in finite samples, so that pointwise inference is again unreliable. The lack of uniformity for pre-testing or post-selection estimators has been widely noted in the classic time-series setting by Shibata (1986); Pötscher (1991); Kabaila (1995); and Leeb and Pötscher (2005).

To remedy this issue, we develop uniformly valid inference in the spirit of Mikusheva (2012); Andrews and Cheng (2012); Andrews, Cheng, and Guggenberger (2011); and Belloni, Chernozhukov, and Hansen (2014) in different contexts, on volatility over a large class of MA($\infty$) models that allow for an asymptotically vanishing noise with a flexible dependence structure. Our inference is thereby more reliable than that of the realized volatility, which simply ignores the impact of small noise when it is difficult to detect. Our inference also allows for model-selection mistakes, which surely occur in the case of an MA($\infty$) data-generating process, and is therefore robust to the dependence structure of noise.

The crux of our uniformity results is that the convergence rate of our estimator depends on various sequences of noise DGPs. Similar to our estimator but in the case of white noise, Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) show that as the variance of the noise vanishes, the convergence rate of the realized kernel estimator improves from $n^{1/4}$ to $n^{1/2}$ with the optimal choice of bandwidth. Their inference, however, is not uniformly valid because there remains a gap between the small noise regime they consider and the no-noise regime. More specifically, they require the noise variance to be greater than $n^{-1}$. This seemingly small gap is not innocuous, because in a finite sample, the noise magnitude may fall into this gap, resulting in a distortion in the prescribed asymptotic distribution. In contrast, Jacod and Protter (2011) (Theorem 16.5.7) study the pre-averaging estimator as the noise variance vanishes at the rate of $n^{-\eta}$ for $\eta \geq 0$. They do not, however, provide a uniformly valid
inference procedure, and the convergence rate of their estimator cannot exceed $n^{1/3}$.\(^1\)

In the case of serially correlated noise, the unknown dependence structure may further plague the efficiency. The pre-averaging estimator by Jacod, Li, and Zheng (2019) and the flat-top realized kernel estimator by Varneskov (2016) converge at the rate of $n^{1/4}$. They do not consider alternative sequences of noise DGPs (particularly those in which noise magnitude and dependence structure interact) that may influence the convergence rate and validity of their inference. We investigate various noise DGPs with infinite-order autocorrelation and arbitrarily shrinking magnitude simultaneously. We show that the convergence rate of our estimator dominates $n^{1/4}$ as noise vanishes, and is determined by the order of noise dependence when noise is sufficiently small. While it is appealing to consider data-driven order selection, such as information criteria, for efficiency gains, in light of the critique by Leeb and Pötscher (2005), we adopt a slightly more conservative order selection procedure than AIC. As such, our inference remains uniformly valid with a slight efficiency loss only in the small-noise regime.

In the case where noise does not vanish, we find that the dependence structure of noise influences the asymptotic variance of our estimator only through the long-run variance of the noise process. In light of this result, using Le Cam’s concept of asymptotic equivalence (Le Cam and Yang (2000)), we establish the minimax optimal efficiency bound in a somewhat restrictive setting where the efficient price is a continuous Itô semimartingale (without leverage effect) and noise follows a Gaussian distribution. The efficiency bound is similar to that of the i.i.d. noise case developed by Reiß (2011), with the variance of the noise therein replaced by its long-run variance. Our estimator reaches the bound when volatility is constant. In general, the efficiency gap between our estimator and the efficiency bound depends on the variability of the volatility.\(^2\)

The literature on the estimation of quadratic variation using noisy high-frequency data is enormous. The earlier works mainly tackle a white microstructure noise.\(^3\) In this paper, we target serially correlated noise using a likelihood-based approach.\(^4\) Hansen, Large, and Lunde (2008) first shed light on the asymptotic equivalence between the maximum likelihood estimator and MA filters. They impose an “essentially white” noise assumption, which requires that conditionally on the latent process the noise is centered and independent, although not necessarily identically distributed. For the sake of presentation, we do not distinguish this type of noise from the white noise, since it is also serially uncorrelated, unlike the “colored” noise setting we consider.

\(^1\)Jacod and Protter (2011) impose an “essentially white” noise assumption, which requires that conditionally on the latent process the noise is centered and independent, although not necessarily identically distributed. For the sake of presentation, we do not distinguish this type of noise from the white noise, since it is also serially uncorrelated, unlike the “colored” noise setting we consider.

\(^2\)The assumptions under which we derive this efficiency bound are unlikely to hold in reality. Yet, they lead to a tractable Gaussian likelihood which facilitates a Le Cam type analysis.

\(^3\)Prominent estimators include, but are not limited to, two-scale or multi-scale estimators by Zhang, Mykland, and Aït-Sahalia (2005) and Zhang (2006); the realized kernel estimator and its extensions by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011); the pre-averaging estimator by Jacod, Li, Mykland, Podolskij, and Vetter (2009) and Jacod, Podolskij, and Vetter (2010); the local method of moments estimator by Reiß (2011); and likelihood-based estimators by Aït-Sahalia, Mykland, and Zhang (2005) and Xiu (2010).

\(^4\)Empirical evidence of autocorrelations in returns of transaction prices goes back to as early as Niederhoffer and Osborne (1966), Simmons (1971), Garbade and Lieber (1977), and Hasbrouck and Ho (1987). Many hypotheses may explain higher-order autocorrelations of returns, such as strategic order splitting (Garbade and Lieber (1977)); optimal control of execution cost (Bertsimas and Lo (1998)); price impact and inventory control (Kyle (1985), Amihud and Mendelson (1980)); the crowd effect or herding (Tóth, Palit, Lillo, and Farmer (2015)); and high-frequency trading.
plement the MA($q$) estimator and demonstrate its desirable performance in extensive simulations with various noise models. Related work that discusses serially dependent noise also include Kahina and Linton (2008); Bandi and Russell (2008); Aït-Sahalia, Mykland, and Zhang (2011); Hautsch and Podolskij (2013); Bibinger, Hautsch, Malec, and Reiß (2019); and Li, Laeven, and Vellekoop (2017). Their assumptions on noise, however, are more restrictive than in our setting.

Our paper is organized as follows. Section 2 sets up the model, Section 3 provides the main results, and Section 4 concludes. The Supplemental Material provides model-selection consistency results using Bayesian information criterion (BIC), (pointwise) inference on estimators of noise autocovariances and autocorrelations, the quadratic representation of our estimator and its connection with realized kernels, Monte Carlo simulations, an empirical study of S&P Composite 1500 index constituents, and all mathematical proofs.

2 Model Setup

We start with notation. For any matrix $A$, $A^\top$ and $A^\dagger$ denote its transpose and Hermitian conjugate, respectively. We denote by $\delta_{i,j}$ the Kronecker delta. The imaginary unit and the indicator function are written as $i$ and $1\{\cdot\}$, respectively. All vectors are column vectors. We write $(a, b, c)$ in place of $(a^\top, b^\top, c^\top)^\top$ for simplicity. $d$-dimensional vectors of 0s and 1s are written as $0_d$ and $1_d$. We use $\|\cdot\|$ to denote the $L^2$ norm. We use $B$ to denote the backward (lag) operator associated with discrete-time time series. We use $K$ as a generic positive constant that may vary from line to line but not depend on $n$. All limits are taken as $n \to \infty$. We use $\overset{L^2}{\to}$ and $\overset{L^s}{\to}$ to denote convergence in law and stable convergence in law, respectively. We write $a_n \sim b_n$ if $a_n \overset{L^2}{\to} b_n$. We write $a_n \lesssim b_n$ if $a_n \overset{L^2}{\to} b_n \lesssim a_n$. We use $a \lor b$ and $a \land b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. We index certain objects below by a superscript $(n)$ to facilitate discussion of uniformity over different sequences of data-generating processes (DGPs).

At each stage $n \geq 1$, transaction prices $\tilde{X}$ are observed at time points $0 = t_0 < t_1 < \ldots < t_{n_T} \leq T$, where $T$ is fixed. Throughout, we assume $n_T$, the number of observations within $[0, T]$, is an observed random variable, whereas $n$ is a non-observable mathematical abstraction. We let $\Delta_n = T/n_T$. We assume $\tilde{X}_{t_i}$ comprises two components:

$$\tilde{X}_{t_i} = X_{t_i} + U_i, \quad 0 \leq i \leq n_T,$$

where $X_{t_i}$ is (the logarithm of) the efficient equilibrium price and $U_i$ is the microstructure noise associated with the $i$th observation.

Specifically, with respect to the efficient price, we assume the following:

**Assumption 1.** The logarithm of the efficient price process $X_t$ is an Itô-semimartingale defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P}^{(n)})$ and satisfies

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + (\delta 1_{|\delta| \leq 1}) * (\mu - \nu)_t + (\delta 1_{|\delta| > 1}) * \mu_t,$$

(2.1)
where $\mu_t$ and $\sigma_t$ are adapted and locally bounded, $W$ is a standard Brownian motion, and $\underline{\mu}$ is a Poisson random measure on $\mathbb{R}_+ \times E$, where $E$ is a Polish space. The compensator $\underline{\nu}$ satisfies $\underline{\nu}(dt, du) = dt \otimes \lambda(du)$ for some $\sigma$-finite measure $\lambda$ on $E$. Moreover, $|\delta(\omega, t, u)| \wedge 1 \leq \Gamma_m(u)$ for all $(\omega, t, u)$ with $t \leq \tau_m(\omega)$, where $\{\tau_m\}$ is a localizing sequence of stopping times and $\{\Gamma_m\}$ a sequence of deterministic functions satisfying $\int \Gamma^r_m(u) \lambda(du) < \infty$ for some $r \in [0, 1)$.

In addition, the process $Z_t = (\mu_t, \sigma_t^2)$ is also an Itô semimartingale on the space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P}^{(n)})$ with the form

$$Z_t = Z_0 + \int_0^t \bar{\mu}_s ds + \int_0^t \bar{\sigma}_s d\tilde{W}_s + \left(\tilde{\delta}_1(\tilde{\delta}_1 \leq 1) \right) \ast (\underline{\mu} - \nu)_t + \left(\tilde{\delta}_1(\tilde{\delta}_1 > 1) \right) \ast \underline{\mu}, \quad (2.2)$$

where $\bar{\mu}_t$ and $\bar{\sigma}_t$ are locally bounded adapted processes, $\tilde{W}$ is a multivariate Brownian motion, potentially correlated with $W$, and $\tilde{\delta}$ is a predictable function such that for some deterministic function $\Gamma_m(u)$, $\|\tilde{\delta}(\omega, t, u)\| \wedge 1 \leq \Gamma_m(u)$ for all $\omega \in \Omega$, $t \leq \tau_m(\omega)$, and $\int \Gamma^2_m(u) \lambda(du) < \infty$.

Assumption 1 allows for the leverage effect and jumps in both the efficient price and its volatility. It accommodates most models in finance and is commonly used to derive in-fill asymptotic results for high-frequency data — for example, Jacod and Protter (2011) and Aït-Sahalia and Jacod (2014), with notable exclusions of long-memory volatility models driven by fractional Brownian motions (Comte and Renault (1996, 1998)).

The parameter of interest is the quadratic variation of $X$ (scaled by $T^{-1}$), which comprises both continuous and discontinuous components:

$$C_T = \frac{1}{T} \int_0^T \sigma_t^2 dt + \frac{1}{T} \sum_{0 \leq t \leq T} (\Delta X_t)^2,$$

where $\Delta X_t = X_t - X_{t-}$. Although estimating the integrated volatility or the jump component of the quadratic variation is of tremendous interest, we do not pursue this agenda in this paper, in which we aim for a practical volatility estimate that depends on as few tuning parameters as possible.\(^5\)

Next, we make an assumption on the arrival of trades:

**Assumption 2.** For each $n \geq 1$, the sequence of observation times $\{t_i : i \geq 0\}$ satisfies $t_0 = 0$ and $t_i = t_{i-1} + \frac{T}{n} \xi_{t_{i-1}} \chi_i$, where the sequence $\{\chi_i : i \geq 1\}$ is i.i.d., $(0, \infty)$-valued, defined on $(\Omega, \mathcal{F}, \mathbb{P}^{(n)})$, and independent of the $\sigma$-field $\mathcal{F}_\infty = \bigvee_{t \geq 0} \mathcal{F}_t$, with $m_j = \mathbb{E}((\chi_i)^j) < \infty$ and $m_1 = 1$, for all $j > 0$. In addition, the process $\xi = (\xi_t)_{t \geq 0}$ is a nonnegative Itô-semimartingale defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P}^{(n)})$ in the form of (2.2), such that neither $\xi_t$ nor $\xi_{t-}$ vanishes.

Assumption 2 allows the arrival rate of transactions to depend on their prices through $\xi_t$. It also accommodates regular sampling, time-changed regular sampling, Poisson sampling, modulated

\(^5\)To achieve robustness to serially correlated noise, prominent nonparametric estimators require three tuning parameters (see two such estimators in our simulation study). Exploring the finer structure of the quadratic variations would require at least one more, rendering these measures impractical to estimate, in particular for illiquid stocks.
Poisson sampling, and predictably modulated random-walk sampling schemes, as discussed in detail by Jacod, Li, and Zheng (2017).\(^6\) We introduce here and below several stochastic processes, e.g., \(\xi_t\) and \(\eta_t\), for which their driving Brownian motions (implicitly defined) are usually different from \(W\) in Assumption 1, but possibly correlated. Note that finding a single Poisson measure that drives the jumps of all processes involved is always possible.

Finally, we assume the microstructure noise is endogenous, heteroscedastic, and serially correlated.

**Assumption 3.** For each \(n \geq 1\), the noise sequence \(\{U_i : i \geq 0\}\) are random variables defined on the probability space \((\Omega, F, \mathbb{P}^{(n)})\). \(\mathbb{P}^{(n)}\) is such that \(\{U_i : i \geq 0\}\) has an MA(\(\infty\)) representation with mean 0:

\[
U_i = \eta_i \iota^{(n)}(B) \varepsilon_i, \quad \text{with} \quad \theta^{(n)}(x) = 1 + \sum_{j=1}^{\infty} \theta_j^{(n)} x^j, \tag{2.3}
\]

where \(\varepsilon_i \sim \mathcal{N}(0, 1)\), defined on \((\Omega, F, \mathbb{P}^{(n)})\), is independent of \(\mathcal{F}_\infty\) and \(\{\chi_i : i \geq 1\}\), and has finite moments of all orders, \((\eta_t)_{t \geq 0}\) is an \(\mathcal{F}_t\)-adapted nonnegative Itô-semimartingale that satisfies the same form of (2.2), and \(\iota^{(n)}\) is a deterministic non-negative number that characterizes the noise magnitude and satisfies \(\iota^{(n)} \leq K\).

Assumption 3 accommodates several empirical features of the microstructure noise. The noise process depends on price \(X\) through \(\eta\), since \(\eta_t\) is \(\mathcal{F}_t\)-adapted, which may be driven by a Brownian motion and a Poisson random measure that are correlated with \(X\). Such dependence is potentially driven by comovement between the price and bid-ask spread or the discreteness of the observed price.\(^7\) That said, this assumption implies zero correlation between any function of the path of \(X\) and \(U_i\) for each \(i\)—the key identifying assumption that separates efficient returns from noise. A fully specified structural microstructure model would be necessary, along with additional observables (e.g., bid-ask prices), if some non-vanishing correlation between \(X\) and \(U\) were allowed for. In this paper, we avoid imposing additional structural assumptions, and instead focus on the reduced-form model of \(\tilde{X}\), while being agnostic about the economic implications of reduced-form parameters, e.g., \(\theta\) and \(\iota^2\). Many structural models yield specific reduced-form ARMA models of returns; for example, Hasbrouck (2007) and, more recently, Andersen, Cebioglu, and Hautsch (2017), with differences only in how the reduced-form parameters relate to structural parameters. Estimating and interpreting

---

\(^6\)Our sampling scheme imposes that conditional on \(\mathcal{F}_{t-1}\), \(t_i\) is independent of \(X_t\) for \(t \geq t_{i-1}\), which we need to derive a desirable central limit theory. This assumption conforms with Assumption O(ii) of Jacod, Li, and Zheng (2017), but is more restrictive than those adopted by Li, Mykland, Renault, Zhang, and Zheng (2014) and Fukasawa (2010), both of which find an asymptotic bias in the CLT of the realized volatility estimator associated with their general sampling scheme (in the absence of microstructure noise). On a related note, Renault and Werker (2011) find that the instantaneous causality relations between price volatility and durations of trades could lead to severely biased volatility estimates.

\(^7\)No one has yet built a perfect model that is meant to be a full describer of reality. Like Jacod, Li, and Zheng (2017) and Jacod, Li, and Zheng (2019), the type of rounding errors our setting allows for is still somewhat restrictive. In the simulation analysis, we investigate a particular form of rounding that appears more realistic but violates Assumption 3, and we show the rounding effect is negligible.
structural parameters in a microstructure model is interesting, yet we leave this direction for future work.

The noise process features flexible serial correlations through its \( \theta^{(n)}(B) \varepsilon \) component, as specified by an MA(\( \infty \)) model. The next assumption spells out restrictions on its spectral density function, 
\[
g(\lambda; \theta^{(n)}) = |\theta^{(n)}(e^{i\lambda})|^2,
\]
such that the sequence of MA processes is uniformly invertible and their long-range serial dependence cannot be arbitrarily strong.

**Assumption 4.** For each \( n \geq 1 \), the spectral density function of \( \theta^{(n)}(B) \varepsilon \) satisfies for some fixed \( \alpha > 3 \),
\[
\inf_{\lambda} g(\lambda; \theta^{(n)}) \geq \frac{1}{K} \quad \text{and} \quad \int_{-\pi}^{\pi} g(\lambda; \theta^{(n)}) e^{i\lambda j} d\lambda \leq K, \quad \forall j \geq 0.
\]

We next introduce our likelihood-based estimator.

3 Main Results

3.1 Likelihood-based Estimation

In contrast to existing nonparametric estimators, we construct a quasi-maximum likelihood estimator (QMLE) in the spirit of White (1982) by posing a misspecified parametric model, for which the likelihood function is available:

\[
dX_t = \sigma dW_t; \quad U_i = i\theta(B)\varepsilon_i, \quad \text{with} \quad \theta(x) = 1 + \sum_{j=1}^{q} \theta_j x^j, \quad \text{and} \quad \varepsilon_i \sim N(0, 1).
\]

In other words, we pretend the efficient price (in logarithm) is a Brownian motion with constant volatility but no drift, and that the noise follows a Gaussian MA(\( q \)) model with the order \( q \) to be determined. Under this model, the observed log-return vector \( Y_n = (Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n_T})^T \), which is defined as

\[
Y_{n,i} = X_{t_i} - X_{t_{i-1}} + U_i - U_{i-1}, \quad 1 \leq i \leq n_T,
\]

follows a reduced-form Gaussian MA(\( q + 1 \)) model. Its \( n_T \times n_T \) covariance matrix \( \Sigma_n \) is given by

\[
\Sigma_n(\sigma^2, \tau^2, \theta) = \sigma^2 \Delta_n I_n + \sum_{h=0}^{q-1} (2\gamma_h - \gamma_{h+1} - \gamma_{h-1}) G_n^h,
\]

where \( (I_n)_{ij} = \delta_{i,j}, \quad (G_n^h)_{ij} = \delta_{h,|i-j|}, \quad \text{and} \quad \gamma_h \) is the \( h \)-th order autocovariance of \( U \):

\[
\gamma_h = \frac{\tau^2}{2\pi} \int_{-\pi}^{\pi} g(\lambda; \theta) e^{i\lambda h} d\lambda, \quad \text{where} \quad g(\lambda; \theta) = |\theta(e^{i\lambda})|^2.
\]

Because \( \theta \) is a nuisance parameter for volatility estimation and is unidentified if \( \iota = 0 \), we reparametrize the likelihood function in terms of strongly identified parameters \( (\sigma^2, \gamma) \), where \( \gamma \) is a
(q + 1)-dimensional vector of noise autocovariances:

\[ L_n(\sigma^2, \gamma) = -\frac{1}{2} \log \det(\Sigma_n(\sigma^2, \gamma)) - \frac{1}{2} \text{tr}(\Sigma_n(\sigma^2, \gamma)^{-1}Y_nY_n^\top), \]  

(3.7)

where, \( \Sigma_n(\sigma^2, \gamma) := \Sigma_n(\sigma^2, \iota^2, \theta) \), and we define \( (\hat{\sigma}^2_n(q), \hat{\gamma}_n(q)) \) as the maximizer of \( L_n(\sigma^2, \gamma) \):

\[ (\hat{\sigma}^2_n(q), \hat{\gamma}_n(q)) = \arg \max_{(\sigma^2, \gamma) \in \Pi_n(q)} L_n(\sigma^2, \gamma). \]  

(3.8)

The parameter space of \((\sigma^2, \gamma)\), denoted by \( \Pi_n(q) \), can be derived from the usual condition that \((\sigma^2, \iota^2, \theta)\) satisfy, i.e., \( \inf_{\lambda} f(\lambda; \gamma) \geq 0 \), where \( f(\lambda; \gamma) = \iota^2 g(\lambda; \theta) \).

Aït-Sahalia and Xiu (2019) show that in the white-noise case, if the noise magnitude is small, the noise variance estimator \( \hat{\sigma}^2_n \) will hit the boundary zero, so that the asymptotic distribution of the volatility estimator, \( \hat{\sigma}^2_n \), becomes non-standard. A similar yet more severe issue occurs here: The estimate \( \hat{\gamma} \) hits the boundary \( \inf_{\lambda} f(\lambda; \gamma) = 0 \) with nontrivial probability.

An easy solution in the white-noise case is to enlarge the parameter space of the nuisance parameter, allowing for negative values of \( \hat{\sigma}^2_n \), so that the asymptotic distribution of \( \hat{\sigma}^2_n \) is not affected by confinement of the parameter space for \( \iota^2 \). We adopt a similar strategy to enlarge the parameter space of \((\sigma^2, \gamma)\) to \( \{(\sigma^2, \gamma) \in \mathbb{R}^{q+2} : \inf_{\lambda} f(\lambda; \sigma^2, \gamma, \Delta_n) \geq 0\} \), where \( f(\lambda; \sigma^2, \gamma, \Delta_n) = \sigma^2 \Delta_n + |1 - e^{i\lambda}|^2 f(\lambda; \gamma) \) is the spectral density of \( Y_n \) under the quasi model. In other words, the parameter space is enlarged such that only the reduced-form MA(\( q + 1 \)) model of observed returns is required to be well defined, and that a well-defined decomposition of observed returns may not exist. On the other hand, the parameter space must be sufficiently “local” to the true value to avoid spurious estimates due to potential use of an overly flexible quasi model (e.g., \( q \) is too large). For this purpose, we choose the following set that imposes constraints on the lower bound of the spectral density function and the decay of autocovariances:

\[ \Pi_n(q) = \left\{ (\sigma^2, \gamma) \in \mathbb{R}^{q+2} : \inf_{\lambda} f(\lambda; \sigma^2, \gamma, \Delta_n) \geq \frac{\Delta_n}{K}, \quad \sigma^2 + |\gamma_0| + \frac{\sum_{j=1}^\infty j^2 |\gamma_j|}{\inf_{\lambda} |\sigma^2 \Delta_n + f(\lambda; \gamma)|} \leq K \right\}. \]  

(3.9)

The likelihood estimator does not have an explicit solution, making it rather cumbersome to analyze. In the case of i.i.d. noise, Xiu (2010) shows the QMLE can be equivalently written as an (iterative) kernel estimator with an implied exponential kernel function and a “bandwidth” that depends on the estimates (and hence is iterative). In the case of an MA(\( q \)) noise, we show in Appendix A.4 that the QMLE is equivalent to a flat-top kernel when noise does not vanish, which explains intuitively why the QMLE works, although it also involves many additional noise parameters.

---

8Note that \( \gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q) \), which is different from how vectors are typically indexed. For convenience, we often treat \( \gamma \) as an infinite dimensional vector, with 0s filled beyond the \((q + 1)\)th entry of \( \gamma \) when no ambiguity exists.

9The constraint (3.9) is essential for proofs. We do not find it critical to impose in our implementation.
3.2 Model Selection

To determine an appropriate order \( q \), we use the Akaike information criterion (AIC), which in our setting can be written as

\[
AIC_n(q) = 2q - 2 \max_{(\sigma^2, \gamma) \in \Pi_n(q)} L_n(\sigma^2, \gamma).
\]

Our choice of order \( q \) will be based on (but not necessarily identical to)

\[
\hat{q}_{n, AIC} = \arg \min_{q \leq n^{1/3}} AIC_n(q).
\] (3.10)

We can also define a similar estimator based on \( \hat{q}_{n, BIC} \), which minimizes \( BIC_n(q) \), defined by replacing the term \( 2q \) in \( AIC_n(q) \) by \( q \log nT \). In the appendix, we prove the model-selection procedure based on BIC is (pointwise) consistent if the true noise-dependence structure follows MA(\( q \)), for some finite \( q \); see Proposition A1 in Appendix A.2. More generally, in Theorem 1 below, we spell out the conditions a desirable order \( \hat{q}_n \) must satisfy, so as to accommodate uniformly valid inference on volatility for a large class of DGPs.

Similar to the case of AR(\( \infty \)) in Shibata (1980), the upper bound on \( q \) precludes MA models with too many parameters for estimation. Asymptotically, this upper bound is not binding, because for all sequences of noise DGPs we consider, \( \hat{q}_{n, AIC} = o_P(n^{1/6}) \) — a claim we prove in Proposition F2 in the appendix. In practice, it appears to be not restrictive either, because for almost all stock-day pairs in our empirical study, the selected orders are smaller than 10 using AIC (or 6 using BIC).

In light of the result on model-selection consistency, we provide (pointwise) inference results on noise parameters \( \gamma \) in Proposition A2. The pointwise asymptotic theory relies on a fixed DGP for the noise process, as well as an unrealistic result of perfect model selection; hence, it provides a misleading picture of the actual finite-sample behavior of the inference. As shown in the classic time-series setting of Leeb and Pötscher (2005), conducting uniformly valid post-selection inference on parameters over a nontrivially large class of DGPs is generally impossible. For volatility estimation in our setting, however, uniformly valid inference is possible for a wide class of DGPs, which we turn to next.

3.3 Uniform Inference on Volatility

Obviously, the class of DGPs cannot be arbitrarily large, so we need restrictions on how the magnitude of the noise and its autocorrelation structure vary with sample size. We denote by \( \kappa^{(n)} \) the \( \infty \)-dimensional vector of autocovariances of \( \theta^{(n)}(B)\varepsilon \), whose components are given by

\[
\kappa_j^{(n)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda; \theta^{(n)}) e^{i\lambda j} d\lambda, \quad j \geq 0.
\] (3.11)

The class of noise models we consider satisfies:
Assumption 5. For any $0 < k < K$ and any sequence $\alpha_n \to \infty$, $q_n^*(k) = o(n^{1/3}(\kappa(n) \lor n^{-1/2})^{4/9})$,

$$\frac{(\kappa(n))^2}{n^{-1} + (\kappa(n))^2} \sum_{j=q_n^*(k)+1}^{\infty} |\kappa_j(n)| = o\left(\frac{q_n^*(k)^{1/2} + \alpha_n}{n^{1/2}} + \frac{\sqrt{\kappa(n)}}{n^{1/4}}\right),$$

where

$$q_n^*(k) := \arg\min_q \left[\frac{kq}{n} + \frac{(\kappa(n))^4}{n^{-2} + (\kappa(n))^4} \sum_{j=q+1}^{\infty} |\kappa_j(n)|^2\right].$$

Intuitively, $q_n^*(k)$ mimics the “oracle” order that AIC selects (up to some constant because of different choices of ks in $(0, K)$), and Assumption 5 effectively requires that this order cannot be too large and that the approximation error induced by selection (the left-hand side) is asymptotically dominated by the estimation error (the right-hand side). Next, we provide two examples to demonstrate that the conditions in Assumption 5 are not restrictive.

**Example 1:** Suppose $n^{1/2} \kappa(n) \to \infty$ and $\theta^{(n)}(B) \varepsilon$ follows an MA($\infty$) model with $|\kappa_j(n)| \sim j^{-\alpha}$ for some $\alpha > 3 \lor 2/\log n$. It is easy to show Assumption 5 holds, because

$$q_n^*(k) \sim n^{1/(2\alpha)} \quad \text{and} \quad \frac{(\kappa(n))^2}{n^{-1} + (\kappa(n))^2} \sum_{j=q_n^*(k)+1}^{\infty} |\kappa_j(n)| \sim n^{-1/2 + 1/(2\alpha)} = o\left(n^{-1/4}(\kappa(n))^{1/2}\right).$$

Jasod, Li, and Zheng (2017) assume $|\kappa_j| \sim j^{-\alpha}$ with $\alpha > 3$. Our condition further sheds light on a trade-off between $\kappa(n)$ and $\alpha$: As $\kappa(n)$ shrinks, $|\kappa_j(n)|$ must decay faster.

**Example 2:** Suppose $\theta^{(n)}(B) \varepsilon$ follows an arbitrary ARMA($p, q$) process with finite $p$ and $q$. Assumption 5 holds because in this case, as long as $\kappa(n) \lesssim 1$ (it can shrink arbitrarily fast),

$$q_n^*(k) \lesssim \log n \quad \text{and} \quad \frac{(\kappa(n))^2}{n^{-1} + (\kappa(n))^2} \sum_{j=q_n^*(k)+1}^{\infty} |\kappa_j(n)| \lesssim n^{-1/2}.$$

Now we are ready to present the main theoretical result, based on which we build uniformly valid inference on volatility:

**Theorem 1.** Let $\{\mathbb{P}^{(n)}\}_{n \geq 1}$ be a sequence of DGPs that satisfies Assumptions 1 - 5. Suppose we select an order $q_n = q_n^{*}\text{AIC} \lor \alpha_n$ with $\alpha_n = O(\log n)$. Then, as $q_n \lor \sqrt{n^{1/2} \kappa(n)} \to \infty$, we have

$$\frac{\tilde{\sigma}^2(q_n) - C_T}{\sqrt{\text{AVAR}(q_n, n)_T}} \xrightarrow{\mathcal{D}} N(0,1),$$

where $\text{AVAR}(q, n)_T$ is given by

$$\text{AVAR}(q, n)_T = \frac{1}{n} \left[ (4q + 6) E(4, \xi)_T + \Delta_{\mu}^{-1/2} \kappa(n) \left(5E(4, \xi)_T C_T^{-1/2} + C_T^{3/2} B(\xi)_T\right) \right], \quad (3.12)$$
where $(\zeta(n))^2$ is the “long-run variance” of the general noise process, given by

$$(\zeta(n))^2 = (\iota(n))^2 g(0, \theta^{(n)}) \int_0^T \frac{\eta_s^2 \xi_s^{-1}}{\xi_s^{-1}} ds,$$

$E(4, \xi)_T$ is a general “quarticity” in the presence of random sampling and jumps, given by

$$E(4, \xi)_T = \frac{1}{T} \int_0^T \frac{\xi_s \sigma_s^4}{\xi_s^{-1}} ds + \frac{1}{T} \sum_{s \leq T} (\Delta X_s)^2 (\xi_s \sigma_s^2 + \xi_{s-} \sigma_{s-}^2),$$

and

$$B(\xi)_T = \frac{2 \int_0^T \eta_s^2 \xi_s^2 ds + \sum_{s \leq T} (\Delta X_s)^2 (\eta_s^2 + \eta_{s-}^2)}{C_T \times \int_0^T \eta_s^2 \xi_s^{-1} ds} + \frac{T \int_0^T \eta_s^2 \xi_s^{-1} ds}{\left( \int_0^T \eta_s^2 \xi_s^{-1} ds \right)^2}.\tag{3.14}$$

Combining with asymptotic variance estimators in Section 3.5, we immediately obtain the following:

**Corollary 1.** Suppose the same assumptions as those in Theorem 1 hold. Let $c_{1-\alpha} = F^{-1}(1-\alpha/2)$, where $F(\cdot)$ is the standard Gaussian cumulative distribution function. We have

$$\lim_{n \to \infty} \Pr(n) (C_T \in CI_n(\alpha)) = 1 - \alpha,$$

where, using $\widehat{\zeta}_n = \sum_{j=-q}^q \widehat{\zeta}_n(\widehat{q}_n)_{|j|}$, the uniformly valid confidence interval is constructed as

$$CI_n(\alpha) = \left[ \widehat{\sigma}_n^2(\widehat{q}_n) + c_{1-\alpha} n^{-1/2} \sqrt{\left( 4 \widehat{q}_n + 6 \right) \widehat{E}_n(4)_T + \frac{\zeta_n^2}{\Delta_n} \left( 5 \widehat{E}_n(4)_T \widehat{\sigma}_n^2(\widehat{q}_n)^{-1/2} + \widehat{\sigma}_n^2(\widehat{q}_n)^{3/2} \widehat{B}_n(\widehat{q}_n)_T \right)} \right].$$

To shed light on the asymptotic behavior of our estimator, we emphasize results for two special DGP sequences:

i. Under $n^{1/2} \iota(n)/(4 \widehat{q}_n + 6) \to \infty$,

$$\text{AVAR}(\widehat{q}_n, n)_T = n^{-1/2} T^{-1/2} \zeta(n) \left( 5 E(4, \xi)_T C_T^{-1/2} + C_T^{3/2} B(\xi)_T \right) + o_p(n^{-1/2} \iota(n)).\tag{3.15}$$

ii. Under $n^{1/2} \iota(n)/(4 \widehat{q}_n + 6) \to 0$,

$$\text{AVAR}(\widehat{q}_n, n)_T = \frac{1}{n} (4 \widehat{q}_n + 6) E(4, \xi)_T + o_p\left( \frac{\widehat{q}_n + 1}{n} \right).\tag{3.16}$$

\[In this case, Theorem 1 requires that $\widehat{q}_n$ approaches $\infty$, so that $4 \widehat{q}_n + 6$ and $\widehat{q}_n$ are in fact of the same order. That said, we prefer this small-sample adjustment that can be established in the case of a finite $\widehat{q}_n$.\]
Case i describes the behavior of our estimator in the presence of "large" noise. The convergence rate is \((\nu(n))^{-1/2}n^{1/4}\), which varies within \([n^{1/4}, n^{1/2}\tilde{q}_n^{-1/2}]\). This result echoes and extends that of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) for the realized kernel estimator whose rate varies within \([n^{1/4}, n^{1/2}]\) in the case of i.i.d. noise. This rate dominates \(n^{1/4}\) — the convergence rate of flat-top realized kernel and pre-averaging estimators by Varneskov (2016) and Jacod, Li, and Zheng (2019).

In the case of small noise (Case ii), the convergence rate is prescribed by \(n^{1/2}\tilde{q}_n^{-1/2}\). When noise is absent, Case ii also shows the efficiency loss compared to the realized volatility estimator is given by a factor \(2\tilde{q}_n + 3\), because realized volatility has knowledge of the absence of noise. Moreover, the bias of the realized volatility estimator is of order \((\nu(n))^{2}n\), which may not vanish in Case ii, because noise is not entirely negligible in this regime.

We now explain our choice of \(\tilde{q}_n\). Recall the noise-dependence structure follows MA(\(\infty\)). Intuitively, a smaller choice of \(\tilde{q}_n\) leads to a more efficient estimator at the risk of a larger bias due to model misspecification \((\tilde{q}_n < \infty)\). In contrast to the somewhat ad-hoc tuning parameters other approaches rely on, our estimate \(\hat{q}_{n,AIC}\) is informative about the minimal order using which the model misspecification bias is negligible. The importance of this guidance on \(q\) is manifested in Case ii, in which the convergence rate clearly improves as \(\hat{q}_n\) decreases.

Nonetheless, instead of fully relying on \(\hat{q}_{n,AIC}\), Theorem 1 requires the use of a certain \(\tilde{q}_n = \hat{q}_{n,AIC} \lor \alpha_n\) that also approaches \(\infty\) slowly if \(n^{1/2}\nu(n)\) is bounded, even when the true model may be of a finite order (and hence \(\hat{q}_{n,AIC}\) is small). Indeed, if the true model is a finite-order MA(\(q\)), we can show that QMLE based on \(\hat{q}_{n,AIC}\) can achieve a convergence rate as fast as \(n^{1/2}\) in Case ii. However, the asymptotic distribution is highly nonstandard because the model-selection bias is of a comparable order to the estimation error. For this reason, we intentionally inflate the order of the employed model, requiring \(\tilde{q}_n \to \infty\), so that a standard asymptotic normal distribution is available in Case ii. \(\alpha_n\) is the single tuning parameter required by our procedure. One possible choice of \(\alpha_n\) is \(\log n_T\), which (potentially) inflates \(\hat{q}_{n,AIC}\) by \(\log n_T\). The choice of \(\tilde{q}_n\) does not affect the asymptotic variance in Case i as (3.15) in fact does not rely on \(\tilde{q}_n\), but it may hurt the efficiency of our estimator in Case ii. As a result, our rate in Case ii is strictly smaller than \(n^{1/2}\) under the conditions in Theorem 1. This efficiency cost is in fact unavoidable for the sake of uniformity, because of the following "impossibility" result in the spirit of Leeb and Pötscher (2008).

To demonstrate this result, we consider a simple setting in which the noise process has no autocorrelation beyond the first lag (so that we use AIC to select \(q\) from \(\{0, 1\}\)), and the noise magnitude \((\nu^{(n)})^2\) is of order \(n^{-1}\) (so that the optimal rate of the volatility estimator is \(n^{1/2}\)). The next proposition shows that even with constant volatility, no uniformly consistent estimator exists for the cumulative distribution function \(G_n(x)\), where

\[
G_n(x) = \mathbb{P}^{(n)} \left( n^{1/2}(\hat{q}^2_{n,AIC} \land 1) - C_T \right) \leq x. 
\]
Proposition 1. For each $x \in \mathbb{R}$, a sequence of DGPs $\{\mathbb{P}^{(n)}\}_{n \geq 1}$ exists that satisfies Assumptions 1 and 2 with $\sigma^2_t = C_T$ for some $C_T$ fixed and all $t \in [0, T]$, and Assumptions 3 and 4 with $n \iota^{(n)} \leq K$ and a single parameter $\theta^{(n)}$ such that

$$\liminf_{n \to \infty} \inf \mathbb{P}^{(n)} \left( |\hat{G}_n(x) - G_n(x)| > \frac{1}{K} \right) > 0,$$

where the infimum extends over all estimators $\hat{G}_n(x)$ of $G_n(x)$.

On a different note, Theorem 1 establishes that our asymptotic distribution is conditionally Gaussian, which is typically the case for estimating the quadratic variation in the presence of noise. Nonetheless, in the absence of noise, the limiting distribution for the realized volatility estimator is a mixture of Gaussian random variables with square root of uniform random variables around the jump times instead of Gaussian, unless the volatility and price processes do not jump together see, e.g., Theorem 5.4.2 of Jacod and Protter (2011). Because of this, the inference procedure is substantially complicated that simulations must be involved in order to achieve a sharp confidence interval, see, page 349 of Aït-Sahalia and Jacod (2014). In the small-noise regime, our asymptotic distribution would run into the same issue if $\hat{q}_n$ is finite. Interestingly, while the requirement on $\hat{q}_n \to \infty$ is motivated from uniformity considerations, we show that this condition also leads to a conditional Gaussian limiting distribution, which facilitates our inference procedure.

Another related point is that the asymptotic distribution of our volatility estimator does not depend on the true distribution of $\varepsilon$. This is not surprising in the case of large noise, see, e.g., Xiu (2010). However, when noise is small, it is possible that certain moments of $\varepsilon$ might affect the asymptotic variance of volatility as the convergence rate improves (just like the cumulant of noise appears in the asymptotic variances of estimators of noise parameters which converge at $n^{1/2}$, see Proposition A2 of the appendix). In this regime, the reason that our volatility estimator has the same asymptotic variance regardless of the true distribution of $\varepsilon$ is again due to that $\hat{q}_n \to \infty$.

3.4 Minimax Efficiency Bound on Volatility Estimation

We now analyze the minimax efficiency bound of volatility estimation from noisy returns in a stylized model in which noise follows MA($q$). Our approach relies on Le Cam’s concept of asymptotic equivalence between experiments. Two sequences of statistical experiments $(\mathcal{E}_n^{(0)}, \mathcal{E}_n^{(1)})$ are asymptotically equivalent if their Le Cam distance $\Delta_{\text{LC}}(\mathcal{E}_n^{(0)}, \mathcal{E}_n^{(1)})$ vanishes asymptotically; see, for example, Le Cam and Yang (2000). Using this approach, Reiß (2011) establishes the minimax efficiency bound of volatility estimation, which is $T^{-3/2} \int_0^T \sigma^2_t dt$ in the case of i.i.d. Gaussian noise. We use his result to establish the minimax bound in our setting. To do so, we need the following assumption, which is motivated from that of Reiß (2011) but we allow for serially correlated noise:

Assumption 6. Suppose $\mathcal{E}_n^{(0)} = \mathcal{E}_n^{(0)}(\alpha, \iota^2, \theta)$ is a statistical experiment generated by observing $\{Y_{n,j}\}_{j=1}^{nT}$ from (3.4), (2.1), and (2.3) under regular sampling interval $T/n$, where $\eta_t = 1$, $\iota^{(n)} = \iota$. 

13
and the distribution of $\varepsilon$ is Gaussian. Furthermore, $X_t$ is continuous, drift $\mu_t$ is zero, and volatility $\sigma^2_t$ is $\alpha$-Hölder continuous, independent of $X_t$, satisfying $\min_{t \in [0,T]} \sigma^2_t \wedge \sigma^{-2}_t \geq K^{-1}$.

In addition, $\mathcal{E}^{(1)}_n = \mathcal{E}^{(1)}_n(\alpha, a^2)$ is another statistical experiment generated by observing $\{Y_{n,j}\}_{j=1}^n$ from (3.4) and (2.1) under regular sampling interval $T/n$, where $U_j$ is i.i.d. centered Gaussian with variance $a^2$. Furthermore, $X_t$ is continuous, drift $\mu_t$ is zero, and volatility $\sigma^2_t$ is $\alpha$-Hölder continuous, independent of $X_t$, satisfying $\min_{t \in [0,T]} \sigma^2_t \wedge \sigma^{-2}_t \geq K^{-1}$.

Assumption 6 imposes independence between $X$ and $\sigma^2$, which we only use to prove Theorem 2 below. This condition is commonly used in the literature to develop the efficiency bound for, for example, volatility estimation by Renault, Sarisoy, and Werker (2017) and jump regressions by Li, Todorov, and Tauchen (2017). Also, Assumption 6 rules out the heteroscedasticity of the noise.

We observe from (3.15) that the asymptotic variance of QMLE with serially correlated noise coincides with that in the white-noise case (Xiu (2010)), except that $\iota^2$ in the latter is replaced by the long-run variance of the dependent noise, namely, $\zeta^2 = \iota^2(1 + \sum_{j=1}^q \theta_j^2)$. Therefore, $\theta$ appears in the asymptotic variance only through $\zeta$. This finding leads to our conjecture that in terms of volatility estimation, $\mathcal{E}^{(0)}_n$ provides the same information as $\mathcal{E}^{(1)}_n$, as long as their noise processes have the same long-run variance. We prove the following theorem:

**Theorem 2.** Suppose Assumption 6 holds. Then, for any $\alpha > \frac{1}{4}$ and $\theta \in \{\theta \in \mathbb{R}^q : \inf_\lambda g(\lambda; \theta) \geq \frac{1}{K}\}$ for some fixed $q$, the experiments $\mathcal{E}^{(0)}_n$ and $\mathcal{E}^{(1)}_n$ with $a^2 = \zeta^2$ are asymptotically equivalent. More precisely, their Le Cam distance satisfies that

$$\Delta_{LC}(\mathcal{E}^{(0)}_n(\alpha, \sigma^2, \iota^2, \theta), \mathcal{E}^{(1)}_n(\alpha, \sigma^2, a^2)) \lesssim n^{-1/4} + n^{1/4 - \alpha} (\log n)^{2\alpha}.$$

Consequently, the minimax efficiency bound for volatility estimation is given by $8T^{-3/2} \zeta \int_0^T \sigma^2_s ds$.

Altmeyer and Bibinger (2015) propose an adaptive estimator that achieves the optimal efficiency in the case of white noise, whereas Jacod and Mykland (2015) impose less restrictive assumptions on noise, allowing for heteroscedasticity but no serial correlations. Their estimator, however, does not achieve the lower bound. In theory, designing a likelihood-based estimator that achieves the efficiency bound is not difficult. More specifically, we first divide $[0, T]$ into $m$ blocks, for which the numbers of observations are (asymptotically) of the same order, then apply QMLE on each block and aggregate these volatility estimates. The resulting estimator will reach the efficiency bound if $m$ diverges slowly. However, the complexity of this procedure deteriorates its finite-sample performance, as shown from simulation results (not included due to space constraints). Moreover, the efficiency gap between the regular QMLE and the efficiency bound depends on the variability of volatility, which, based on our calibration with real data, is rather small. We therefore recommend using our simpler estimator in practice.
In this section, we develop pre-averaging-based estimators of asymptotic variances. We need two sequences of integers $k_n$ and $k'_n$, satisfying $k_n \sim n^{2/3}, k'_n \sim n^{7/8}$, and a non-zero real-valued function $g : \mathbb{R} \to \mathbb{R}$, supported on $[0,1]$, which is continuous and piecewise $C^1$ with a piecewise Lipschitz derivative $g'$ and $g(0) = g(1) = 0$. We also adopt a truncation strategy (Mancini (2001)) to handle jump-related quantities, for which we define:

$v_n = \alpha(k_n \Delta_n)^{\varpi}$, for some $\alpha > 0, \varpi \in (0,1/2)$.

We construct the estimator of $E(4,\xi)_T$ in (3.13) as $\hat{E}_n(4)_T = \tilde{C}_n(4)_T + \hat{D}_n(4)_T$ using the pre-averaging approach:

$$\tilde{C}_n(4)_T = \frac{1}{Tk_n^2 \Delta_n} \left( \int_0^1 g(s)^2 ds \right)^2 \sum_{m=1}^{n_T - 2k_n} \left( \bar{Y}(g)^n_m \right)^2 \left( \bar{Y}(g)^n_{m+k_n} \right)^2 \mathbb{1}_{\{ |\bar{Y}(g)^n_m| \leq v_n, |\bar{Y}(g)^n_{m+k_n}| \leq v_n \}},$$

$$\hat{D}_n(4)_T = \frac{1}{Tk_n f_0^1 g(s)^2 ds} \sum_{m=k_n,1}^{n_T - k_n - k'_n} \left( \bar{Y}(g)^n_m \right)^2 \mathbb{1}_{\{ |\bar{Y}(g)^n_m| > v_n \}} \left( \bar{c}(g)^n_m + \bar{c}(g)^n_{m-k'_n} \right),$$

where pre-averaged returns and spot volatilities are given by, respectively,

$$\bar{Y}(g)^n_i = \sum_{j=1}^{k_n-1} g \left( \frac{j}{k_n} \right) Y_{n,i+j}, \quad \bar{c}(g)^n_i = \frac{1}{k'_nk_n \Delta_n} \int_0^1 g(s)^2 ds \sum_{m=1}^{k'_n} \left( \bar{Y}(g)^n_{i+m} \right)^2 \mathbb{1}_{\{ |\bar{Y}(g)^n_{i+m}| \leq v_n \}}.$$

These estimators are the same as those constructed by Aït-Sahalia and Xiu (2016) for i.i.d. noises. Despite their low convergence rate, these estimators are also consistent in this more general setting, because of the choice of a large local window size $k_n$ that averages out the impact of the noise. Because of the jump truncation, Assumption 1 imposes that $r < 1$, which is necessary for the consistency proof.

Finally, we provide the estimator of $B(\xi)_T$ in (3.14) using

$$\hat{B}_n(\bar{q})_T = \left| \frac{1}{\bar{\sigma}^2(\bar{q})} \left( \frac{1}{\tilde{\gamma}(\bar{q})_0 - \tilde{\gamma}(\bar{q})_1} \right) \times (\hat{B}'_n(1) + \hat{B}'_n(2)) + \left( \frac{1}{\tilde{\gamma}(\bar{q})_0 - \tilde{\gamma}(\bar{q})_1} \right)^2 \times \hat{B}'_n(3) \right| \land n_T,$$

where, with $\bar{Y}(g)^n_m$ and $\bar{c}(g)^n_m$ defined in (3.19),

$$\hat{B}'_n(1) = \frac{1}{n_T} \sum_{m=1}^{n_T-k_n-k'_n} (Y_{n,m})^2 \bar{c}(g)^n_m,$$

$$\hat{B}'_n(2) = \frac{1}{2Tk_n} \int_0^1 g(s)^2 ds \sum_{m=1}^{n_T-k'_n} ((Y_{n,m})^2 + (Y_{n,m+k'_n})^2) \left( \bar{Y}(g)^n_m \right)^2 \mathbb{1}_{\{ |\bar{Y}(g)^n_m| > v_n \}},$$

$$\hat{B}'_n(3) = \frac{1}{2Tk_n} \int_0^1 g(s)^2 ds \sum_{m=1}^{n_T-k'_n} ((Y_{n,m})^2 + (Y_{n,m+k'_n})^2) \left( \bar{Y}(g)^n_m \right)^2 \mathbb{1}_{\{ |\bar{Y}(g)^n_m| > v_n \}}.$$
\[ \hat{B}_n'(3) = \frac{1}{4nT} \sum_{m=1}^{nT-k'_n} (Y_{n,m})^2 (Y_{n,m+k'_n})^2. \]

We demonstrate the consistency of \( \hat{B}_n(q)_T \) in Appendix F.6.

**3.6 Implementation**

We discuss the implementation of QMLE in this section. Apparently, directly calculating the inverse of \( \Sigma_n(\sigma^2, \gamma) \) would be computationally expensive when evaluating the likelihood function at each stage of an optimization routine. To avoid this problem, the classic time-series literature adopts an approximation approach of Whittle (1951). Unfortunately, in Appendix A.3, we show the Whittle estimator is inconsistent in our in-fill asymptotic setting, even if the noise is i.i.d. Gaussian and the efficient price is a Brownian motion with constant volatility (hence, our QMLE is in fact the MLE).

We instead implement exact likelihood through the state-space representation of an MA model. To avoid the issue of weakly identified parameters, our implementation leverages an auxiliary reduced-form MA\((q + 1)\) model of the observed noisy returns:

\[
Y_{n,i} = \phi(B) \epsilon_i, \quad \text{with} \quad \phi(x) = 1 + \sum_{j=1}^{q+1} \phi_j x^j, \quad 1 \leq i \leq n, \quad \epsilon \sim \mathcal{N}(0, \chi^2).
\]

(3.21)

**Algorithm 1.** Our algorithm starts as follows:

1. Select the optimal order, \( \hat{q}_{n,\text{AIC}} \), of the MA process (3.21) for \( Y_n \) using AIC, defined by (3.10) but rewritten in terms of \( \chi^2 \) and \( \phi \).

2. Obtain exact quasi-likelihood estimates of \( \hat{\chi}^2 \) and \( \hat{\phi}_j \) for \( 1 \leq j \leq \hat{q}_n + 1 \), using the state-space representation of (3.21) and Kalman filtering (see, e.g., Gardner, Harvey, and Phillips (1980)),\(^{11}\) where \( \hat{q}_n = \hat{q}_{n,\text{AIC}} \lor \log nT.\(^{12}\)

3. Construct volatility and noise autocovariance estimators using the above estimates:

\[
\hat{\sigma}^2_n(\hat{q}_n) = \Delta^{-1}_n \hat{\chi}^2 \left( 1 + \sum_{j=1}^{\hat{q}_n+1} \hat{\phi}_j \right)^2,
\]

\[
\hat{\gamma}_n(\hat{q}_n)_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{\chi}^2 e^{ijk\lambda} \left( 1 + \sum_{l=1}^{\hat{q}_n+1} \hat{\phi}_l e^{i\lambda l} \right)^2 d\lambda, \quad 0 \leq j \leq \hat{q}_n,
\]

which are obtained by comparing different parameterizations of the return autocovariances.

\(^{11}\)Packages of standard programming software (e.g., R and Matlab) are available that implement a likelihood estimator for MA models, despite the fact that some packages rely on Whittle approximations. Our codes are also available upon request.

\(^{12}\)In our empirical analysis of S&P 1500 index components, see Appendix C, almost all selected orders are smaller than 10 using AIC.
4. Solve \( q_n + 1 \) nonlinear equations for \( q_n + 1 \) model parameters \((\hat{\iota}_n^2(q_n), \hat{\theta}_n(q_n))\) from \( \hat{\gamma}_n(q_n) \) obtained in Step 3:

\[
\hat{\gamma}_n(q_n)_j = \hat{\iota}_n^2(q_n) \sum_{l=0}^{\hat{q}_n-j} \hat{\theta}_n(q_n)_{l+j}, \quad 0 \leq j \leq \hat{q}_n. 
\]

(3.22)

A Newton-Raphson algorithm that converges quadratically is available from Wilson (1969).

Effectively, Step 4 is to find \( q_n + 1 \) model parameters of the MA(\( q_n \)) noise process from up-to-\( q_n \)-th order autocovariances \( \hat{\gamma}_n(q_n)_j \), \( 0 \leq j \leq \hat{q}_n \). This practice is common in the classic time-series analysis. For instance, Box, Jenkins, and Reinsel (2007) recommend using this algorithm to find initial values based on autocovariances for the maximum likelihood estimation of an MA model.

Note that Step 3 is sufficient for volatility and noise autocovariance estimation, which is rather simple. If one is further interested in \((\iota^2, \theta)\), a unique solution \((\hat{\iota}_n^2(q_n), \hat{\theta}_n(q_n))\) exists from Step 4, with probability approaching 1 when noise is sufficiently large relative to sample size. When noise is small, however, these parameters are weakly identified, and (3.22) may have no solution such that \( \hat{\iota}_n^2(q) \) is positive and \( \hat{\theta}_n(q) \) is real. Studying the inference of \((\iota^2, \theta)\) in the small-noise case might be interesting, but the primary objective of this paper is uniformly valid inference on volatility, whose inference is not affected by Step 4.

3.7 Consistency of Noise Autocovariances and Autocorrelations

Recall that in (3.8) and Step 2 of Algorithm 1, we have defined and implemented estimators of noise autocovariances. We further propose estimators of noise autocorrelations. The \( \infty \)-dimensional vectors of autocovariances and autocorrelations of \( U \) under \( \mathbb{P}^{(n)} \) can be written as

\[
\gamma_j^{(n)} = (\iota^{(n)})^{2T} \int_0^T \eta_s^2 \xi_{s}^{-1} ds \times \kappa_j^{(n)}, \quad j \geq 0, \quad \text{and} \quad \rho_j^{(n)} = \kappa_j^{(n)}/\kappa_0^{(n)}, \quad j \geq 1, 
\]

where \( \kappa_j^{(n)} \) is given by (3.11).

We define \( \hat{\rho}_n(q_n) \) as follows. If (3.22) has a solution such that \( \hat{\iota}_n^2(q_n) \) is positive and \( \hat{\theta}_n(q_n) \) is real, we let\(^{13}\)

\[
\hat{\rho}_n(q_n)_j = \frac{\hat{\gamma}_n(q_n)_j}{\hat{\gamma}_n(q_n)_0}, \quad j \geq 1.
\]

Otherwise, we let

\[
\hat{\rho}_n(q_n) = 0.
\]

In light of their definitions, we can regard these estimators as “hard-thresholding” estimators in that higher-order autocovariances and autocorrelations estimates are truncated to zero beyond the selected order \( q_n \).

\(^{13}\)Estimates of autocovariances and autocorrelations are, of course, zero beyond the \( q_n \)-th lag.
We now present the uniform consistency results of \( \hat{\gamma}_n(\hat{q}_n) \) and \( \hat{\rho}_n(\hat{q}_n) \) with respect to \( \gamma^{(n)} \) and \( \rho^{(n)} \) under \( L^2 \)-norm, where all vectors are regarded as \( \infty \)-dimensional. The consistency of \( \hat{\gamma}_n(\hat{q}_n) \) is needed for Corollary 1. For completeness, we provide the (pointwise) central limit results on noise parameters in the supplementary appendix.

**Theorem 3.** Let \( \{p^{(n)}\}_{n \geq 1} \) be a sequence of DGPs that satisfies Assumptions 1 - 5. If we select the order \( \hat{q}_n \) that satisfies \( \hat{q}_n = \hat{q}_n,_{\text{AIC}} \vee \alpha_n \) with \( \alpha_n = O(\log n) \), it holds that

\[
\|\hat{\gamma}_n(\hat{q}_n) - \gamma^{(n)}\| = o_P(n^{-1} \vee (\iota^{(n)})^2).
\]

If, in addition, we assume \( P^{(n)}(\iota^{(n)} \leq \Delta_n^{1/2}) \to 0 \), it holds that

\[
\|\hat{\rho}_n(\hat{q}_n) - \rho^{(n)}\| = o_P(1).
\]

While consistent estimation of autocorrelations requires more restrictive DGPs, Theorem 3 allows for small and vanishing noises. The small-noise case is a relevant scenario in practice, as shown from our empirical study in the supplementary appendix. However, Jacod, Li, and Zheng (2017) do not allow for this. As shown from our simulation results in the appendix, our estimators outperform theirs by a wide margin, especially in this small-noise regime.

**4 Conclusion**

We propose a simple volatility estimator based on the likelihood of an MA model, whose order is selected based on AIC (or BIC). We establish uniformly valid inference on volatility over a large and flexible class of noise DGPs, featuring autocorrelations of an infinite order and an arbitrarily vanishing noise magnitude. The convergence rate of our estimator is greater than or equal to \( n^{1/4} \), which depends on the noise magnitude and its dependence structure. Our estimator requires one single tuning parameter in order selection, and it always guarantees the positivity of volatility estimates. For these reasons, it delivers more desirable finite-sample performance than alternative nonparametric estimators, as our simulations in the supplementary appendix show. Our empirical study of S&P 1500 stocks in the appendix highlights the limitations of applying the realized volatility estimator to a large cross section of stocks — no safe frequency exists that one can use without accounting for the microstructure noise.

Important byproducts of our approach are the estimates of noise autocovariances and autocorrelations. These estimates are potentially informative about structural parameters of certain microstructure models, which we leave for future work. Allowing for small noise, our approach resembles a threshold estimator, which gives zero autocovariance estimates beyond the lag selected by the information criteria. This feature delivers superior performance in the finite sample, particularly when noise is relatively small. Empirically, we find that autocovariances of observed returns in recent years last for a much shorter period of time than in earlier years. This finding indicates market efficiency has improved substantially, potentially due to the popularity of electronic and algorithmic
trading. In a cross-sectional comparison, the autocovariances of small-cap stocks tend to persist for a longer period than the large caps, perhaps due to limits to arbitrage or for liquidity reasons.

References


