Sankhyā B : The Indian Journal of Statistics
https://doi.org/10.1007/s13571-022-00280-7

Review of Statistical Approaches for Modeling High-Frequency Trading Data

Chiranjit Dutta and Nalini Ravishanker
University of Connecticut, Storrs, USA
Kara Karpman
Middlebury College, Middlebury, USA
Sumanta Basu
Cornell University, Ithaca, USA

Abstract

Due to technological advancements over the last two decades, algorithmic trading strategies are now widely used in financial markets. In turn, these strategies have generated high-frequency (HF) data sets, which provide information at an extremely fine scale and are useful for understanding market behaviors, dynamics, and microstructures. In this paper, we discuss how information flow impacts the behavior of high-frequency (HF) traders and how certain high-frequency trading (HFT) strategies significantly impact market dynamics (e.g., asset prices). The paper also reviews several statistical modeling approaches for analyzing HFT data. We discuss four popular approaches for handling HFT data: (i) aggregating data into regularly spaced bins and then applying regular time series models, (ii) modeling jumps in price processes, (iii) point process approaches for modeling the occurrence of events of interest, and (iv) modeling sequences of inter-event durations. We discuss two methods for defining events, one based on the asset price, and the other based on both price and volume of the asset. We construct durations based on these two definitions, and apply models to tick-by-tick data for assets traded on the New York Stock Exchange (NYSE). We discuss some open challenges arising in HFT data analysis including some empirical analysis, and also review applications of HFT data in finance and economics, outlining several research directions.

AMS (2000) *subject classification.* Primary 62M10, 62M05; Secondary 62P20.
Keywords and phrases. Asynchronous data, durations, GARCH models, high-frequency trading data, jumps, volatility.

1 Introduction

With advances in automation, algorithmic trading has largely replaced floor-based trading in many financial markets. These automated trading

platforms allow for unprecedented speed of order placement and execution, thereby creating financial data sets that are of increasingly higher frequency (Cont, 2011). While a small firm may have traded tens of times each day in the 1990s, high-frequency trading (HFT) firms today can execute several thousand trades each day.

The advent of high-frequency trading (HFT) has triggered substantial interest among financial traders and policymakers alike, due to HFT's promise of offering deeper insights into the workings of financial markets. Information contained in high-frequency financial data has been shown to predict market movements such as realized volatility and certain higher-order moments of returns distributions (Easley et al., 2021). In addition, this data is useful for studying market microstructure, which examines the process by which assets are traded and the resulting consequences (O'Hara, 1997). For example, access to high-frequency (HF) data sets can provide greater insight into price discovery mechanisms (Tay et al., 2004) and the presence of informed traders (Easley et al., 2012b). All of these aspects are valuable for regulators and policymakers.

Tick-by-tick (high-frequency) data sets are more information-rich than asset price data collected at regular intervals, e.g., monthly, daily, or even intra-day. This is because—in addition to the trajectory of the price process—these high-frequency data sets provide us with the *number of trades executed* for a particular stock within a small time interval. Trade volumes provide additional information about how attractive a particular stock is to investors. Thus, analyzing the number of events that occur in a certain time window—or equivalently, the duration between trades or other events—is at the heart of statistical modeling of HFT data and is precisely what makes analysis of HFT data both unique and challenging.

Since trades can occur at any point in time, HFT data on asset prices are usually *irregularly spaced time series*. These data sets also exhibit some stylized features that contribute to new modeling challenges. For instance, HFT data exhibit intra-day diurnal effects, i.e., higher activity during the start and close of the market and less activity during the middle of the trading day. Moreover, most variables used in HFT data analyses (e.g., volumes and bid-ask spreads) are non-negative and exhibit strong temporal dependence, which introduces non-trivial challenges in modeling and estimation.

Statistical modeling of HFT data broadly falls into three categories. The first approach aggregates HFT data into regularly spaced bins (1 min, 5 min, etc.) and then applies models to the regular time series. Such models include the autoregressive conditional heteroscedastic (ARCH) model,

the generalized ARCH (GARCH) model, and the stochastic volatility (SV) model (Tsay, 2005). The second approach builds on the *point process analysis* literature to model the occurrence of events of interest (e.g., a single trade of an asset). The third approach models *sequences of durations* between consecutive occurrences of events. Both the second and third families have tailored strategies to incorporate other relevant information into the modeling framework, e.g., information about the relative occurrence of buys versus sells, as well as the direction and magnitude of price changes.

In this paper, we focus primarily on the third approach: duration models. We describe two common methods for defining events of interest and thereby constructing inter-event durations: (i) using a pre-specified threshold on price changes in the asset, and (ii) explicitly including information on the asset's order flow. We review existing statistical methods for duration modeling and contrast the two event definition methods through the lens of a popular class of conditional duration models, which we apply on HFT data sets.

The rest of the paper is organized as follows. Section 2 provides examples of how HF traders use available information to determine when to place and change orders. Section 3 details some of the stylized features of HFT data. Section 4 briefly reviews the models for regularly spaced time series and discusses modeling event times using point processes. Section 5 discusses some of the existing open challenges in HFT data. Section 6 describes econometric applications where HFT data are routinely used. Section 7 presents a summary and discussion.

2 Background on High-Frequency Trading (HFT)

Technological innovations, combined with a series of regulatory changes in the early 2000s, ushered in the era of high-frequency trading. HFT is characterized by both high speeds and high volumes: traders operate on the order of micro- or nanoseconds, which they achieve through a combination of sophisticated technology and co-location, the practice of placing computers in the same area as an exchange's servers. To cover these fixed costs—and since HF traders earn very small profits per trade—firms place large volumes of orders. These firms aim to generate profits by exploiting small amounts of predictive power on large quantities of orders (Easley et al., 2012a).

Most HFT firms use limit orders, meaning the trader sets a price above (below) which she is unwilling to (buy) sell. These limit prices are referred to as the bid and ask prices, respectively. If a counterparty is not immediately

willing to take the opposite side of the trade, then the order is placed in a limit order book. Many exchanges govern their books using automated rule-based order-matching. Most often this means that orders are executed first according to price priority and then according to time priority. Limit orders can remain on the book for a substantial period of time before the limit price is reached and the trade executed; however, traders may modify or cancel existing orders as part of their trading strategies. Hence the order book is highly dynamic.

HFT firms implement these strategies using automated algorithms that introduce, adjust, and cancel orders as new information is received. Importantly, these algorithms do not operate in clock time, placing an order every minute, for example. Instead the algorithms speed up or slow down their rate of activity based on information they glean from external sources (e.g., macroeconomic events, company earnings announcements) and from other market participants. Consider the HFT strategy known as pinging. Pinging involves placing small immediate-or-cancel limit orders: if liquidity is available at the limit price or better, the order is executed immediately; otherwise the order is canceled. To understand why this might be useful, suppose that firm A places a large buy order and firm B detects this liquidity using immediate-or-cancel limit orders. Firm B can then frontrun firm A's buy order and sell the shares back to firm A at a higher price, thereby generating a profit.

Other HFT strategies include spoofing (placing a large number of orders on one side of the book, with the intention of moving the security's price), redirecting liquidity to the firm's own dark pool, and performing statistical arbitrage (taking advantage of short-lived price differences across trading venues or between related securities). See Goldstein et al. (2014) for further discussion. A significant number of trades are block trades, i.e., high-volume trades that are typically defined as 10,000 or more shares. Large trades may cause a significant price impact, hence block trades are often negotiated through experienced intermediaries who can fragment the order and/or search for counterparties before sending the trade for execution (Keim and Madhavan, 1996; Hasbrouck, 2007).

Since they can place orders at ultra low latencies, HF traders are able to swiftly react to information events. When a trader believes that the price of an asset will rise (fall), she may want to buy (sell) that asset in large quantities and as quickly as possible in order to maximize her short-term profits. Thus greater trading intensity, or equivalently shorter durations between transactions, can signal that an information event has occurred and that there are informed traders in the market. Indeed Easley and O'Hara

(1992) develops a theoretical microstructure model in which longer durations are correlated with the absence of information events. Dufour and Engle (2000) and Manganelli (2005) provide empirical evidence confirming these propositions.

3 Stylized Features of HFT Data

The availability of HFT data has opened new avenues for empirical studies of market microstructure. Before embarking on such analyses, it is important to understand the underlying features of such data that are not observed at lower frequencies. These unique characteristics of HFT data are referred to as *stylized facts* and are usually formulated in terms of qualitative properties of asset returns and durations. These properties are commonly observed across many financial assets, financial markets, and intra-day time periods. In this section, we review some stylized features of HFT data that are discussed in the literature.

Irregular Spacing High-frequency financial transactions typically occur at irregularly spaced time points as trading takes place. Therefore the time durations between consecutive data points are not the same, leading to *irregular time series*. In any given time interval, transactions for a given stock may arrive rapidly, separated by short durations, or arrive slowly, with longer durations between arrival times (see Fig. 1, top and bottom left). Empirical evidence shows that market activity tends to peak around the market opening and closing times, and exhibits varying patterns of intra-day and intra-week behavior (Yan and Zivot, 2003).

Diurnal Effect Intra-day transactions, and hence durations, often exhibit a periodic pattern. For instance, consider inter-event durations, such as trade durations, mid-quote (change) durations, price durations, volume durations, excess volume durations, or excess depth durations, all of which have been well studied in the literature. Such durations often exhibit a *diurnal effect*, which refers to high trading intensity (shorter durations) during the opening and closing periods of the trading day, with relatively lower trading activity (longer durations) around noon (see Fig. 1, top right) (Engle and Russell, 1998). There are many approaches discussed in literature for adjusting durations for the diurnal effect (Tsay, 2005).

Non-negativity and Temporal Dependence Most variables related to HFT data are non-negative (e.g., prices, volumes). Moreover, many

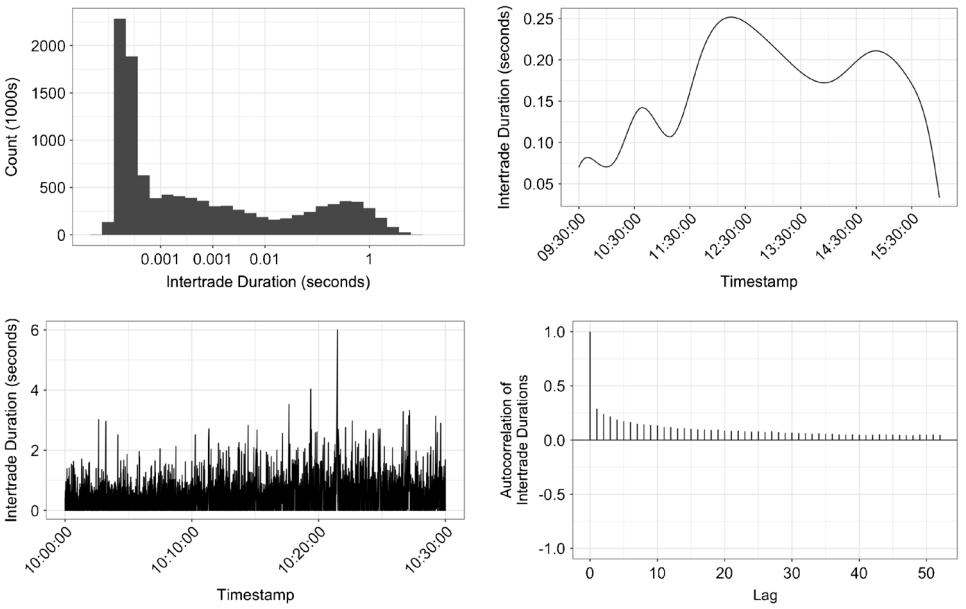

Figure 1: Features of high-frequency financial transactions, illustrated on Bank of America data. **Top left**: histogram of inter-trade durations for the first quarter of 2018. The median time between trades is 0.000118 seconds, with minimum and maximum durations of 0.000005 and 19.963947 seconds, respectively. **Top right**: cubic spline fitted to durations over the course of a single trading day. Notice the increased activity (shorter durations) near the market opening and closing. **Bottom left**: durations between 10:00 and 10:30 AM on a single trading day. Financial durations tend to be clustered, with short (long) durations following other short (long) durations. **Bottom right**: sample autocorrelation function of the duration series

variables, including volumes, spreads, and market depth are positively autocorrelated, i.e., high (low) values tend to be followed by high (low) values. Let ρ_k be the autocorrelation between the k lags of a stationary time series X_t, with mean μ. The autocorrelation is written as $\rho_k = \frac{\gamma(k)}{\gamma(0)}$, where $\gamma(k)$ $= \text{cov}(X_{t+k}, X_t) = \text{E}[(X_{t+k} - \mu)(X_t - \mu)]$. Often X_t exhibits strong persistence; that is, the temporal dependence is non-zero over a long range of lags, with the autocorrelations, ρ_k, decaying in hyperbolic fashion rather than in exponential decay: $\lim_{k\to\infty} \rho_k / ck^{-\alpha} = 1$, with $\alpha \in (0, 1)$ and $c > 0$ (Beran, 1994) (see Fig. 1, bottom right).

Asynchronous Trading Asynchronous (also referred to as non-synchronous) trading of multiple assets is an important stylized feature of HFT data (Fan et al., 2012). In general, different stocks have different trading frequencies, and a single stock's trading intensity often varies intra-day and intra-week. Since the exact timing of transactions within any two stocks are likely to occur independently, we do not expect the trading to be synchronous. Non-synchronous data typically occur due to trading effects (e.g., some assets do not trade at certain periods of the day) or timing effects (e.g., trading happens in different time zones). Suppose there are two stocks A and B and that they are independent, that A trades more frequently than B, and that B stops trading two hours before the market closes. Also suppose that a news item arrives just before closing time on a Monday. Stock A is more likely to show the effect of the news on the same day, while stock B may react with a day's lag. This will in turn differentiate the autocorrelations of the returns of A and B and affect the nature of their cross-correlation function. These features become relevant when we consider models for multivariate HFT time series. The Epps effect has been discussed as the decreasing estimated correlation between two stocks when the sampling frequency increases primarily due to asynchronicity of price observations and possible lead-lag relationships between asset prices (Epps, 1979; Renò, 2003). Hayashi et al. (2005) discuss the bias in estimates of the covolatility based on non-synchronous data.

Jumps in Price Processes The presence of jumps in price processes is one of the stylized features of financial data, and can occur due to specific macroeconomic events (Evans, 2011) or some unexpected news announcements. In empirical studies, it has been shown that the decomposition of daily variation into its continuous and jump components can better explain the volatility dynamics (Andersen et al., 2007; Aït-sahalia et al., 2012; Song et al., 2021). Many statistical tests have been developed to detect jumps from discretely observed prices, see Jiang and Oomen (2005), Barndorff-Nielsen and Shephard (2006), Lee and Mykland (2008), and Aït-sahalia et al. (2012).

Modeling jumps is related to addressing sudden and relatively large changes observed in real stock prices and the implied volatility smile phenomenon (Cont and Tankov, 2004).

4 Approaches for Modeling HFT Data

As mentioned in Section 1, there are several approaches for modeling HFT data. One way to accommodate HFT data in volatility modeling is by

aggregating tick-by-tick data into regularly-spaced bins (e.g., 1 min, 5 min) and then applying models for regularly spaced time series (Tsay, 2005), as discussed in Section 4.1. Although this is a useful approach, it is not generally preferred since aggregation can induce loss of information. Furthermore, traditional time series models of volatility do not account for the intra-day periodicity exhibited by return volatility, i.e., for the systematic patterns that occur over the course of a trading day and that may lead to model misspecification, as illustrated by Andersen and Bollerslev (1997). The classical volatility models are discussed and followed by their extensions to include realized measures of volatility, see So et al. (2021) for an excellent review of univariate and multivariate volatility models for HFT data.

4.1. Regularly Spaced Time Series Models for Aggregated Data Let $\{y_t\}_{t=1}^n$ denote a real-valued, discrete-time stochastic process, and let $\mathcal{F}_t$ be the information set up to time t. In most financial applications, y_t is the log-return of an asset at time t, defined as $y_t = \ln P_t - \ln P_{t-1}$, where P_t is the price of the asset at time t, and ln denotes natural logarithm. The conditional variance of y_t is $h_t = \mathrm{Var}(y_t|\mathcal{F}_{t-1})$ and the conditional mean of y_t is $\mu_t = \mathrm{E}(y_t|\mathcal{F}_{t-1})$. The volatility at time t is given by $h_t^{1/2}$.

GARCH and its Extensions The seminal work of Engle (1982) on the autoregressive conditional heteroscedastic (ARCH) model laid the foundations for modeling the heteroscedasticity (changing variance) of the log-returns process. Later, Bollerslev (1986) proposed a more flexible model, an extension of ARCH known as generalized ARCH (GARCH). The key feature of both models is the ability to accommodate volatility clustering, a key stylized fact of financial time series.

Let y_t denote the (mean-subtracted) returns at time t. Then y_t is said to follow a GARCH(P, Q) model if

$$\begin{aligned} y_t &= h_t^{1/2} z_t, \\ h_t &= \alpha_0 + \sum_{i=1}^{P} \alpha_i y_{t-i}^2 + \sum_{j=1}^{Q} \beta_j h_{t-j}, \quad t = 1, \ldots, T, \end{aligned} \tag{4.1}$$

where $\{z_t\}$ is a sequence of i.i.d. random variables with zero mean and unit variance. In Eq. 4.1, P refers to the lag orders of the returns series while Q refers to the lag orders of the conditional variance series. Sufficient conditions

for the conditional variance h_t to be positive are $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i = 1, \ldots, P$, and $\beta_j \geq 0$ for $j = 1, \ldots, Q$. The constraint $\sum_{i=1}^{max(P,Q)}(\alpha_i + \beta_i) < 1$ ensures stationarity. For $Q = 0$, this process reduces to the ARCH(P) model.

There have been numerous extensions of the GARCH model that are important for modeling the stylized features of financial returns. One such improvement is to use a different distribution for the random variable, z_t, in order to capture the distribution's heavier tails (e.g., we may use Student t-distribution with low degrees of freedom). To capture the asymmetry phenomenon in volatility and to overcome a restrictive model specification, the exponential GARCH (EGARCH) model was proposed by Nelson (1991). Detailed discussions on GARCH modeling and its extensions are provided in Tsay (2005). GARCH models have also been used extensively in the context of high frequency data. Hansen and Lunde (2005), Hansen et al. (2003), and Andersen and Bollerslev (1998) are useful references and the `R` packages *rugarch* by Ghalanos (2020) and *MSGARCH* by Ardia et al. (2019) contains several useful functions that allow users to fit different GARCH models.

Stochastic Volatility Models Stochastic volatility (SV) models were developed to model the heteroscedasticity of y_t by considering a latent (unobserved) stochastic process for volatility, see Taylor (1982, 1994), and Harvey and Shephard (1996). In particular, the evolution of the logarithm of y_t's conditional variance is described by a stochastic process whose dynamics are assumed to be autoregressive. The basic form of SV model is the following:

$$\begin{aligned} y_t = h_t^{1/2}\epsilon_t, \quad \ln h_{t+1} = \alpha + \phi \ln h_t + \eta_t, \\ (\epsilon_t, \eta_t) \sim \mathcal{N}(0, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\eta^2, \end{pmatrix} \quad t = 1, \ldots, T, \end{aligned} \tag{4.2}$$

where $\ln h_t$ is assumed to follow a stationary AR(1) process with $|\phi| < 1$. The latent process h_t can be interpreted as a random flow of information in financial markets and ϕ is the persistence in the volatility. The error terms, ϵ_t and η_t, are independent Gaussian white noise sequences. A fundamental difference between GARCH and SV models is that, in the GARCH framework, given the information set $\mathcal{F}_{t-1}$, the time-varying volatility is assumed to follow a deterministic, rather than stochastic, evolution.

The usefulness of using SV models lies in the fact that they provide greater flexibility in describing stylized facts. The volatility asymmetry, which refers to the different impacts of positive and negative shocks of equal

magnitude on volatility, has been addressed in Jacquier et al. (2004), who proposed an alternate specification of Σ as:

$$(\epsilon_t, \eta_t) \sim \mathcal{N}(0, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & \rho\sigma_\eta \\ \rho\sigma_\eta & \sigma_\eta^2, \end{pmatrix} \quad t = 1, \ldots, T. \tag{4.3}$$

If the correlation ρ is negative, then a decrease in ϵ_t will be associated with an increase in η_t and hence associated with higher contemporaneous and subsequent volatilities through h_t. This will allow us to model the leverage effect, which is the association of negative return with an increase in volatility (Black, 1976). Most financial time series exhibit kurtosis higher than the one resulting by incorporating conditional heteroscedasticity into a normal process. This has been studied in Chib et al. (2002), Harvey et al. (1994), and Jacquier et al. (2004), where ϵ_t in Eq. 4.2 is allowed to follow a Student t-distribution. Detailed extensions of the SV models including its multivariate extensions are discussed in Yu and Meyer (2006), Chib et al. (2009), among others. The `R` package *stochvol* (Hosszejni and Kastner, 2019) contains several functions that allow users to fit different stochastic volatility (SV) models estimated using Bayesian methods. Amongst many others, Barndorff-Nielsen and Shephard (2002), Takahashi et al. (2009), and Stroud and Johannes (2014) are useful references for the applications of stochastic volatility models for high-frequency data.

Realized GARCH Models With the availability of HF data, numerous realized measures of volatility has been introduced in literature including realized variance, bipower variation, realized range and many others (Andersen and Bollerslev 1998; Barndorff-Nielsen and Shephard 2002, 2004; Martens and Van Dijk 2007). Realized measures are more informative than squared returns and hence found more useful for modeling and forecasting future volatility. Andersen and Bollerslev (1998) found realized volatility to be a better measure of volatility than the measures based on daily returns since the former measure provides noise reduction and more temporal stability.

Let P_t be the observed price of an asset for day $t = 1, \ldots, n$. Divide each day into fixed M sub-intervals and let $\Delta = \frac{1}{M}$ denote the length of time between the two consecutive observations. To obtain realized variances at 5-min sampling frequency, set $\Delta = 300$ (seconds) and similarly for 1-min sampling frequency, set $\Delta = 60$ (seconds). Let $\{P_{t-1+j\Delta}\}_{j=1}^{M}$ denote the sequence of observed prices for a day t at the sampling frequency $\Delta = \frac{1}{M}$. In case of missing observed prices, previous tick method or linear interpolation

between adjacent ticks (Zivot and Wang, 2007) can be used to estimate the missing observations. The jth intraday return for day t is defined as

$$y_{j,t} = \ln P_{t-1+j\Delta} - \ln P_{(t-1)+(j-1)\Delta}, \quad j = 1, \ldots, M. \tag{4.4}$$

RV_t (realized variance for day t) is defined as the cumulative squared sum of the intraday returns $y_{j,t}$

$$RV_t = \sum_{j=1}^{M} y_{j,t}^2; \tag{4.5}$$

$\sqrt{RV_t}$ is known as the realized volatility.

Hansen et al. (2012) proposed Realized GARCH for the joint modeling of returns and realized measures of volatility. This model can also be referred to as a GARCH model that makes use of realized measures. The main idea is a measurement equation given by Eq. 4.6 for RV_t that relates the realized measure (e.g., realized variance) to the conditional variance of returns and facilitates modeling of the dependence between returns and future volatility. The realized GARCH model for simultaneous modeling of returns and realized variance is

$$\begin{aligned} y_t &= h_t^{1/2} z_t, \\ h_t &= \alpha_0 + \beta_1 h_{t-1} + \gamma RV_{t-1}, \\ RV_t &= \xi_0 + \xi_1 h_t + \tau(z_t) + u_t, \end{aligned} \tag{4.6}$$

where h_t is the volatility of the asset at time t, $z_t \sim$ i.i.d.$(0,1)$ and $u_t \sim$ i.i.d.$(0, \sigma_u^2)$, with z_t and u_t being mutually independent. $\tau(z_t)$ is known as the leverage function that captures the leverage effect, the dependence between returns and future volatility. The empirical results in Hansen et al. (2012) shows that Realized GARCH outperforms GARCH when applied to Dow Jones Industrial Average stocks and an exchange traded index fund.

In the literature, there have been several extensions and applications of Realized GARCH. Gerlach and Wang (2016) extended the class of Realized GARCH models to include more efficient realized measures such as realized range (RR). So and Xu (2013) proposed to model the intraday returns $y_{j,t}$ in Eq. 4.4 and intraday volatility using a GARCH-RV model with realized volatility to forecast intraday Value-at-Risk (VaR) and intraday volatility.

Realized Stochastic Volatility (SV) Models Takahashi et al. (2009) extended the well-known Stochastic volatility models to jointly model daily

returns and the realized volatility. The proposed model also takes into account the bias in the realized volatility induced by the presence of non-trading hours and market microstructure noise in transaction prices. The specification of the joint model is

$$\begin{aligned} y_t &= \exp(h_t/2)\epsilon_t, \\ h_{t+1} &= \mu + \phi(h_t - \mu) + \eta_t, \\ RV_t &= \xi + h_t + u_t, \\ h_1 &= \mu + \eta_0, \eta_0 \sim N\left(0, \frac{\sigma_\eta^2}{1-\phi^2}\right), \\ & \begin{pmatrix} \epsilon_t \\ \eta_t \\ \rho_t \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho\sigma_\eta & 0 \\ \rho\sigma_\eta & \sigma_\eta^2 & 0 \\ 0 & 0 & \sigma_u^2 \end{pmatrix}\right). \end{aligned} \tag{4.7}$$

In the equation for RV_t, the inclusion of the constant term ξ and the noise term corrects the bias due to microstructure noise and non-trading hours. The parameter ρ captures the correlation between the returns y_t and the future volatility h_{t+1}. The above is an extension of the SV model in Eq. 4.3 to include realized measures.

Asai et al. (2017) proposed a new model to take into account both asymmetry and long memory in which the unobserved time series of log-volatility $\ln h_t$ follows an AutoRegressive Fractionally Integrated Moving Average (ARFIMA) process similar to the approach by Shirota et al. (2014) and the observed series of returns follows the stochastic volatility model with a heavy-tailed distribution. Extreme value distributions have also been incorporated in realized stochastic volatility models by employing the generalized hyperbolic skew Student's t-distribution (Takahashi et al., 2016).

4.2. Modeling Jumps in Price Process The analysis of jumps (instantaneous and discrete moves) in asset price processes is well discussed in the literature; see Cont and Tankov (2004), Carr et al. (2002), Vasileios (2015) and references therein. There are two broad categories of financial models with jumps. The first category consists of jump-diffusion models, where a diffusion process captures normal asset price variations while the Poisson-driven jump part captures large market movements. Different jump-diffusion models arise depending upon the distributional assumptions. The second category consists of jump models that allow for infinitely many jumps in finite time intervals. Examples of this category of models include the variance

gamma model (Madan and Seneta, 1990; Carr et al., 1998), the hyperbolic model (Eberlein and Keller, 1995), the CGMY model (Carr et al., 2002), and the finite moment log stable process (Carr and Wu, 2003).

Models with finite jumps Let $X = (X_t)_{t\geq 0}$ be the logarithmic price of a financial asset that is defined on a filtered probability space $(\omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Here $\mathcal{F}_t$ represents the information available at time t, $t \geq 0$. Following the model setup in Christensen et al. (2014), we assume that X operates in an arbitrage-free frictionless market, implying that X belongs to the class of semimartingale process. We assume that X can be represented by a jump-diffusion model to allow for stochastic volatility and finite price jumps as

$$X_t = X_0 + \int_0^t a_s ds + \int_0^t \sigma_s dW_s + \sum_{i=1}^{N_t^J} J_i \quad t \geq 0, \tag{4.8}$$

where X_t is the log price at time t, $a = (a_t)_{t\geq 0}$ is a locally bounded and predictable drift term, $\sigma = (\sigma_t)_{t\geq 0}$ is an adapted càdlàg volatility process, and $W = (W_t)_{t\geq 0}$ is a standard Brownian motion. Here $N^J = (N_t^J)_{t\geq 0}$ is a counting process representing the total number of jumps in X and $J = (J_i)_{i=1,\ldots,N_t^J}$ is a sequence of nonzero random variables denoting the corresponding jump sizes.

The model specified in Eq. 4.8 nests many popular continuous models as special cases such as the geometric Brownian motion with an Ornstein-Uhlenbeck process for log-volatility (Alizadeh et al., 2002), the stochastic volatility model with log- normal jumps generated by a non-homogeneous Poisson process (Andersen et al., 2002), and the affine class of models (Duffie et al., 2000). In practice, we observe discretely sampled observations which at an individual level can contain substantial noise. One source of noise is from microstructure effects that arise due to bid-ask spreads, or price discreteness (Black, 1986). The other source is due to the presence of outliers which can be attributed to delayed trade reporting on block trades, fat-finger errors, bugs in the data feed, misprints, decimal misplacement, incorrect ordering of data, etc. (Christensen et al., 2014). The model setup for the tick level data that takes into account the sources of noise and restricting to the unit time interval $t \in [0, 1]$ can be specified as

$$Y_{i/N} = X_{i/N} + u_{i/N} + O_{i/N} \quad i = 1, \ldots, N, \tag{4.9}$$

where u is i.i.d. noise process such that $\mathrm{E}(u) = 0$, $\mathrm{E}(u^2) = w^2$ and u and X are independent(Christensen et al., 2014). Here, $O_{i/N} = \mathbb{I}_{i/N\in\mathcal{A}_N}\mathcal{S}_i$, where $\mathcal{A}_N$ is a random set containing the number of occurrences of outliers and

their sizes are given by $(\mathcal{S}_i)_{i=1,\ldots,N_1^0}$. Assuming $\mathcal{A}_N$ to be a.s. finite, model it by

$$\mathcal{A}_N = \left\{ \frac{[NT_i]}{N} : 0 \leq T_i \leq 1 \right\}, \tag{4.10}$$

where $(T_i)_{i=1,\ldots,N_1^0}$ are the arrival times of another counting process $N^O = (N_t^O)_{t\geq 0}$. We assume that O is mutually independent of X and u, which implies that N^J is independent of N^O meaning that the counting process generating jumps and counting process generating outliers are independent of each other. The model in Eq. 4.9 can be considered as an extension to the standard microstructure component formulation considered in the high frequency finance literature (Barndorff-Nielsen and Shephard, 2005) which also accommodates the outlier process.

Models with Infinite Jumps Recently, several studies based on real asset returns have suggested the use of infinite activity models.

Consider a Lévy process with infinite jump activity and microstructure noise, which is considered as one of the simplest models for high frequency financial data. Following the model setup in Wang et al. (2021), consider a one-dimensional Lévy process $X = \{X_t\}_{t\geq 0}$ defined on some probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ over a fixed time horizon $t \in [0, T]$. The model can be represented as

$$X_t = \mu t + \sigma W_t + J_t, \tag{4.11}$$

where $\mu \in \mathcal{R}$ and $\sigma \in [0, \infty)$ are respectively the drift and the variance parameters, $W = \{W_t\}_{t\geq 0}$ is a Wiener process, and $J = \{J_t\}_{t\geq 0}$ is an independent pure-jump Lévy process. The jump component J is defined as

$$\begin{aligned} J_t &= J_{1t} + \tilde{J}_{2t}, \quad J_{1t} = \int_0^t \int_{|x|>1} x\mu(dx, ds), \\ \tilde{J}_{2t} &= \int_0^t \int_{0<|x|\leq 1} x(\mu(dx, ds) - \nu(dx)ds), \end{aligned} \tag{4.12}$$

where μ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}\backslash 0$ with mean measure $\nu(dx)dt$ such that $\int_{\mathbb{R}\backslash 0}(|x|^2 \wedge 1)\nu(dx) < \infty$. Typically, X_t represents the log-return or log-price process $\log(S_t/S_0)$ of an asset with price process $\{S_t\}_{t\geq 0}$. In this case, the parameter σ is called the constant volatility of the process and contributes to the total "variability" of the process X. The model framework (4.11)–(4.12) can also be used to accommodate the observations of the

process which may be contaminated by random errors. Let us assume that the observations take the form

$$Y_{t_j} = X_{t_j} + \epsilon_{t_j} \quad j = 0, \ldots, n, \tag{4.13}$$

with equally-spaced discrete times $0 = t_0 < t_1 \cdots < t_n = T$ such that $t_j - t_{j-1} = T/n$. The assumption of constant volatility is quite restrictive and can be relaxed to stochastic volatility when the microstructure noise can be ignored. It is generally believed that when using medium range frequencies such as 5-min or daily observations, the microstructure noise is negligible. For such data, we may consider the model

$$dX_t = \beta_t dt + \sigma_t dW_t + dJ_t, \quad t \in [0, T], \tag{4.14}$$

where $W = \{W_t\}_{t \geq 0}$ is a Weiner process, $J = \{J_t\}_{t \geq 0}$ is a suitable pure-jump semimartingale, and $\beta = \{\beta_t\}_{t \geq 0}$ and $\sigma = \{\sigma_t\}_{t \geq 0}$ are càdlàg adapted processes. A more detailed discussion about the statistical inference for processes with infinitely many jumps can be found in (Mies et al., 2020).

The `R` package *yuima* (Brouste et al., 2014) has a comprehensive framework for the simulation and inference of stochastic differential equations and other stochastic processes. It allows one to specify stochastic differential equations of abstract types, including one- or multidimensional diffusion processes driven by a Wiener process or a fractional Brownian motion with general Hurst parameter, with or without jumps (i.e. driven by Lévy processes). Iacus and Yoshida (2018) provide more details on the usage of *yuima* package.

Jump Testing Identification and testing of price jumps in asset prices are very important in the context of financial and economic activities such as portfolio re-balancing and risk management, Barndorff-Nielsen and Shephard (2006) proposed a non-parametric test for the existence of jumps based on the ratio of bipower variation and realized quadratic variation, while Lee and Mykland (2008) proposed a non-parametric procedure to detect the exact timing of jumps at the intra-day level using the ratio of realized return to estimated instantaneous volatility. Jiang and Oomen (2008) used an approach based on "swap variance" (accumulated difference between the simple return and log return) to detect the presence of jumps. Sen (2009) proposed a non-parametric method for detecting jumps using the functional data analysis (FDA) technique, which requires no assumptions from the functional volatility process beyond smoothness and integrability. Aït-Sahalia and Jacod (2009) compared two higher order realized power variations with

different sampling intervals to develop a test statistic for the null hypothesis of no jumps. More recently, a rank jump test has been proposed by (Li et al., 2019). In this procedure, the jump matrix is tested for its rank at simultaneous jump events in market returns, as well as in individual assets. A more detailed review of jump tests can be found in (Mukherjee et al., 2020; Tsai and Shackleton, 2016; Bjursell and Gentle, 2012)

In R, the *highfrequency* package (Boudt et al., 2021a), provides functionality and a comprehensive framework to manage high-frequency data. It helps to clean and match high-frequency trades and quotes data, calculate various liquidity measures, estimate and forecast volatility, detect price jumps and investigate microstructure noise and intraday periodicity. The function *AJjumpTest* implements the jump test proposed by Aït-Sahalia and Jacod (2009) while the function *BNSjumpTest* examines the procedure based on Barndorff-Nielsen and Shephard (2006). The function *intradayJumpTest* can be used to test jumps using the theory of Lee and Mykland (2008) and *JOjumpTest* examines the jumps based on Jiang and Oomen (2008). The function *rankJumpTest* implements the rank jump test of Li et al. (2019).

4.3. Modeling Event Times with Point Processes Let t denote the calendar time and let $t_i,\ i = 1, 2, \ldots$, be the random time of occurrence for the ith event. We assume $0 \leq t_i < t_{i+1}$, thereby excluding the possibility of simultaneous occurrence of events. The sequence $\{t_i\}$ is a simple point process in $\mathbb{R}$. If we also observe a mark, W_i, then we refer to $\{t_i, W_i,\ i = 1, 2, \ldots\}$ as a *marked* point process. Marks can indicate different types of events, such as the arrival of buys, sells or certain limit orders. The ith duration is a non-negative random variable defined as the time interval between the events occurring at times t_{i-1} and t_i: $x_i = t_i - t_{i-1}$, for $i = 1, 2, \ldots$, with $t_0 = 0$. A counting process $N(t)$ associated with $\{t_i\}$ is defined by $N(t) = \sum_{i \geq 1} \mathbb{1}[t_i \leq t]$, which is a right-continuous step function with upward jumps of magnitude one at each t_i. These three views of high-frequency data (the point process, durations, and the counting process) are often interchangeable, allowing for rich modeling scenarios and interpretations. How well we can model and predict will of course depend on the quality and format of the data, which in turn depend on underlying institutional settings and recording systems (Harris, 2003; Hautsch, 2011).

Hautsch (2011) discusses four broad approaches for modeling a point process as a function of its history, $\mathcal{F}_t$: (i) models for intensity processes, (ii) models for hazard processes, (iii) models for duration processes, and (iv) models for counting processes. Exogenous predictors, if available, may

be used as well. There is a distinction between discrete-time models and continuous-time models.

The intensity is the instantaneous arrival rate of an event at time t, conditional on the history, $\mathcal{F}_t$ (Daley and Vere-Jones, 2003):

$$\lambda(t; \mathcal{F}_t) = \lim_{\Delta t \to 0} P(1 \quad \text{event} \quad \text{in } [t, t + \Delta t] \mid \mathcal{F}_t),$$

and may be modeled by an autoregressive conditional intensity model (Engle and Russell, 1998). A general class of intensity models can be defined, as in Hautsch (2011), by

$$\lambda(t; \mathcal{F}_t) = \lambda_0(t; g_2(\mathcal{F}_t)) g_1(t; \mathcal{F}_t), \tag{4.15}$$

where $\lambda_0(.)$ denotes a baseline intensity, and $g_1(.)$ and $g_2(.)$ enable us to capture dependence on time-varying covariates or past history. Various choices of λ_0, g_1 and g_2 give rise to well-known models, including the proportional intensity, proportional hazards, and accelerated failure time models. These models can be extended to accommodate long range dependence (Hautsch et al., 2006), for multivariate marked counting process modeling (Russell, 1999), or for stochastic conditional intensity (SCI) models (Bauwens and Hautsch, 2006). Hautsch (2011) discusses dynamic versions of intensity models, defined in continuous time, for univariate and multivariate cases. These include the autoregressive conditional intensity (ACI) model—which is a dynamic extension of the proportional intensity (PI) model (Russell, 1999)—as well as models based on linear self-exciting Hawkes processes (Hawkes, 1971), where the intensity is governed by the sum of negative exponential functions of time to all previous events (Swishchuk and Huffman, 2020). The `R` package *hawkes* by Zaatour (2014) can be used to simulate Hawkes processes both in univariate and multivariate settings. The package contains functions that can be used to compute various moments of the number of jumps on a given interval, separated by a lag.

Given the sequence of inter-event durations, $\{x_i\}$, where x_i has p.d.f. $f(x_i; \mathcal{F}_{i-1})$ and survival function $S(x_i; \mathcal{F}_{i-1})$, the hazard function, $h(x_i; \mathcal{F}_{i-1})$, is defined as the conditional instantaneous risk that the ith event happens in a small time interval $(x_i, x_i + \Delta]$,

$$h(x_i; \mathcal{F}_{i-1}) = f(x_i; \mathcal{F}_{i-1}) / S(x_i; \mathcal{F}_{i-1}).$$

There are close connections between intensity models and hazard models. As mentioned earlier, by setting specific forms for the terms in Eq. 4.15, we can obtain the proportional intensity (PI), proportional hazards (PH), or accelerated failure time (AFT) models. Kalbfleisch and Prentice (2011), Kleinbaum

and Klein (2010), Cox and Oakes (2018), Cox (1972) provide excellent discussions of several hazard process models, including the Cox proportional hazards model and accelerated failure time models. These models have been applied in a variety of settings. For example, the duration of unemployment is studied in a wide range of theoretical and empirical econometric papers, including Lancaster (1979) and Heckman and Singer (1984). Lane et al. (1986) also demonstrates a novel application of the proportional hazards model to the study of bank failure. The R package *survival* by Therneau (2021) contains several functions that allow users to fit these models.

The count process model is obtained by specifying the joint distribution of the number of points in equally spaced intervals of length Δ. A simple count data model is given by

$$N_j^{\Delta} | \mathbf{z}_j^{\Delta} \sim Po(\lambda),$$

where $\mathbf{z}_j^{\Delta}$ is a vector of covariates associated with N_j^{Δ} and N_j^{Δ} is defined as the number of events in the interval $[j\Delta, (j+1)\Delta]$ for $j = 1, 2, \ldots$. Fleming and Harrington (2011) provides an excellent introduction to counting processes. Cameron and Trivedi (2013) discusses more general models by using the negative-binomial distribution and double Poisson distribution introduced by Efron (1986). Dynamic count approaches are also used to model the behaviour of discrete, positive-valued time series, such as bid-ask spreads (the magnitude of trade-to-trade price changes). Let $\{y_i\}_{i=1}^n$ be a time series of counts. Autoregressive conditional Poisson (ACP) models are useful for modeling dynamic intensity processes based on aggregated data. These models can be defined as

$$y_i | \mathcal{F}_{i-1} \sim Poi(\lambda_i), \ \lambda_i = \omega + \sum_{j=1}^{P} \alpha_j y_{i-j} + \sum_{j=1}^{Q} \beta_j \lambda_{i-j}. \tag{4.16}$$

This specification was proposed by Rydberg and Shephard (2000) and its extensions are considered by Heinen (2003). Detailed extensions of the dynamic models for discrete data and its multivariate extensions are discussed in Hautsch (2011). The R packages *acp* by Vasileios (2015) and *tscount* by Liboschik et al. (2017) can be used to fit Autoregressive Conditional Poisson models.

4.4. Modeling Durations Conditional on Past Information Engle and Russell (1998) introduces a general class of time series models for positive-valued random variables, referred to as multiplicative error models (MEM). MEM express the dynamics of the variables of interest (e.g., durations, volume, volatility) as the product of the expectation of the process—conditional

on the available information—and an i.i.d. positive-valued error term with unit mean. This model specification parallels the GARCH specification. The autoregressive conditional duration (ACD) model was the first univariate MEM model introduced in high-frequency econometrics to model the dynamic behavior of the random times between trades (Engle and Russell, 1998). This model was generalized to any non-negative valued process by Engle (2002b).

Let $\{D_i\}_{i=1}^n$ be a discrete time series of raw inter-event durations, defined as $D_i = t_i - t_{i-1}$. Let $\{x_i\}_{i=1}^n$ be a discrete time series of adjusted inter-event durations on $(0,\infty)$, where $x_i = \frac{D_i}{f(t_i)}$, $f(t_i)$ is a deterministic function (Tsay (2005)) consisting of the cyclical component of t_i. Here n is the number of events. Let $\mathcal{F}_{i-1}$ be the information set available at the $(i-1)$th event and let the conditional mean function be $\psi_i = E(x_i|\mathcal{F}_{i-1})$. This notation is used throughout Section 4.4.

Autoregressive Conditional Duration Models (ACD) The ACD model proposed by Engle and Russell (1998) is one of the most popular methods for modeling inter-event durations and acts as a benchmark model for further developments of conditional duration models. We say that $\{x_i\}_{i=1}^n$ follows an ACD(p,q) model if it can be expressed as

$$x_i = \psi_i \epsilon_i, \tag{4.17}$$

where, conditional on $\mathcal{F}_{i-1}$, ψ_i can dynamically evolve as

$$\psi_i = \omega + \sum_{j=1}^{p} \alpha_j x_{i-j} + \sum_{j=1}^{q} \beta_j \psi_{i-j}. \tag{4.18}$$

The ACD model specification assumes that the standardized inter-event durations ϵ_i follow an i.i.d. process defined on positive support and that $E(\epsilon_i) = 1$. The conditions $\omega > 0, \alpha_j \geq 0$ for $j = 1,\ldots,p$, and $\beta_j \geq 0$ for $j = 1,\ldots,q$ ensure that the conditional inter-event durations are positive. By letting $\eta_i = x_i - \psi_i$ (a martingale difference sequence by construction), the ACD(p,q) model can be formulated as an ARMA$(\max(p,q),q)$ model for x_i as

$$x_i = \omega + \sum_{j=1}^{max(p,q)} (\alpha_j + \beta_j) x_{i-j} - \sum_{j=1}^{q} \beta_j \eta_{i-j} + \eta_i. \tag{4.19}$$

Hence a sufficient condition for x_i to be covariance-stationary is given by $\sum_{j=1}^{p} \alpha_j + \sum_{j=1}^{q} \beta_j < 1$ (Pacurar, 2008).

Engle and Russell (1998) discusses parameter estimation using the method of maximum likelihood. Many variations of ACD model specifications have been used in the literature with the assumption that the innovations follow exponential, Weibull, gamma and generalized gamma distributions. Engle and Russell (1998) also discusses the application of WACD and EACD models to IBM transactions data, where WACD and EACD refers to the ACD model with Weibull and exponential innovations, respectively.

Log ACD Models The model specification of ACD(p, q) is quite restrictive since it requires non-negativity constraints on the model parameters to ensure positive-valued inter-event durations. Log ACD_1 and Log ACD_2 models were introduced by Bauwens and Giot (2000) to (i) provide a more flexible structure without any sign restrictions on model parameters, and (ii) to facilitate the inclusion of exogenous variables in the model. The general Log $\text{ACD}_1(p, q)$ model can be written as

$$\begin{aligned} x_i &= e^{\psi_i}\epsilon_i, \\ \psi_i &= \omega + \sum_{j=1}^{p} \alpha_j \ln x_{i-j} + \sum_{j=1}^{q} \beta_j \psi_{i-j}, \end{aligned} \tag{4.20}$$

where $\sum_{j=1}^{max(p,q)}(\alpha_j + \beta_j) < 1$ is required for weak stationarity of x_i. Similarly, for the model specification of Log $\text{ACD}_2(p, q)$, ψ_i takes the form

$$\psi_i = \omega + \sum_{j=1}^{p} \alpha_j \frac{x_{i-j}}{e^{\psi_{i-j}}} + \sum_{j=1}^{q} \beta_j \psi_{i-j}, \tag{4.21}$$

where $\sum_{j=1}^{q} \beta_j < 1$ is required to ensure weak stationarity of x_i. Here ϵ_i are assumed to be an i.i.d. process defined on positive support with $E(\epsilon_i) = 1$. Both specifications of the Log ACD model are more flexible but retain many of the characteristics of the ACD model. Bauwens and Giot (2000) applies the Log ACD_2 model to price durations, relative to the bid-ask quote process, of three securities listed on the New York Stock Exchange: IBM, DISNEY, and BOEING. It also investigates the influence of certain characteristics of the trade process (e.g., trading intensity, average volume per trade, and average spread) on the bid-ask quote process.

Other Types of ACD Models Fernandes and Grammig (2006) proposes a new family of ACD models known as augmented autoregressive conditional duration (AACD) models, which allow asymmetric responses to small and

large shocks. This is a generalization of the ACD process that applies a Box-Cox transformation with parameter $\lambda \geq 0$ to the conditional inter-event duration process, ψ_i. The AACD model can be written as

$$\begin{aligned} x_i &= \psi_i \epsilon_i, \\ \psi_i^{\lambda} &= \omega + \alpha \psi_{i-1}^{\lambda} [|\epsilon_{i-1} - b| + c(\epsilon_{i-1} - b)]^{\nu} + \beta \psi_{i-1}^{\lambda}, \end{aligned} \tag{4.22}$$

where $\omega > 0$, $\alpha > 0$, and $\beta > 0$ and ϵ_i are assumed to be an i.i.d. process defined on positive support. The shocks impact curve is given by $g(\epsilon_i) = [|\epsilon_i - b| + c(\epsilon_i - b)]^{\nu}$. The shift parameter, b, helps identify the asymmetric response implied by the shocks impact curve, while the rotation parameter, c, determines the clockwise ($c < 0$) and anti-clockwise ($c > 0$) rotation. The shape parameter, ν, induces concavity ($\nu \leq 1$) or convexity ($\nu \geq 1$) in the shocks impact curve. AACD models nest many existing ACD models, such as the ACD model of Engle and Russell (1998) and both specifications of the Log ACD models from Bauwens and Giot (2000).

The Stochastic Conditional Duration (SCD) model was introduced by Bauwens and Veredas (2004) under the assumption that a dynamic stochastic latent variable governs the evolution of inter-event durations. The general model specification of SCD(p, q), as in Thavaneswaran et al. (2015), is given by:

$$\begin{aligned} x_i &= e^{\psi_i} \epsilon_i, \\ \psi_i &= \omega + \sum_{j=1}^{p} \alpha_j x_{i-j} + \sum_{j=1}^{q} \beta_j \psi_{i-j} + z_i, \end{aligned} \tag{4.23}$$

where $z_i|\mathcal{F}_{i-1}$ are i.i.d. $N(0, \sigma_z^2)$ variables, $\epsilon_i|\mathcal{F}_{i-1}$ follows a distribution with positive support, and the z_i's and $\epsilon_j|\mathcal{F}_{i-1}$ are independently distributed for all i, j. This model is a generalized version of the model proposed in Bauwens and Veredas (2004). One of the most important features of the latent variable is that it helps to capture the random flow of information, which in the case of financial markets is very difficult to observe directly but drives the duration process. This model is a counterpart to the stochastic volatility model introduced by Taylor (1982).

The major difference between an ACD process and an SCD process is that SCD processes are doubly stochastic processes. The conditional expected duration in the case of SCD models is a random variable, while for ACD models, it is a fixed function of unknown parameters. Parameter estimation in SCD models is a difficult task since it requires integrating out the latent variable. Bauwens and Veredas (2004) estimates the model parameters based on the quasi-maximum likelihood technique using the Kalman

filter. In empirical studies, SCD models have been applied to various inter-event durations (e.g., trade durations, price durations and volume durations) of BOEING, COCA COLA, DISNEY, and EXXON stocks (Bauwens and Veredas (2004)). The authors compare the SCD model with the Log ACD model and find that the SCD model provides a superior fit when the Weibull distribution is used for innovations.

Extreme Value Modeling of Durations Modeling extreme events is popular in finance for quantifying risk, stock market shocks, and large fluctuations in financial data. Embrechts et al. (2013) is an excellent reference outlining the theory on the subject of extreme events modeling. Rocco (2014) provides an extensive survey of the distributional assumptions for calculating different risk measures (e.g., Value-at-Risk, Expected Shortfall) and for the study of dependence and contagion across markets under stress.

Block trades are one example of how extreme value theory is used in modeling durations. Block trades occur when there is a high-volume transaction in a security that is privately negotiated and traded outside of the open market. Typically block trades consist of at least 10,000 shares of a stock or $100,000 of bonds. Zheng et al. (2016) shows that duration sequences of such block trades may exhibit both heavy tails and extreme values; thus it proposes modeling such durations using the Fréchet ACD model, i.e., the usual ACD model with Fréchet innovations. The Fréchet distribution is a special case of a generalized extreme value distribution, and has a heavier right tail than other non-negative distributions such as the gamma and Weibull distributions. Zheng et al. (2016) analyzes durations of block trades from the Hong Kong Stock Exchange (SEHK) and the London Stock Exchange (LSE), demonstrating a better fit using the Fréchet ACD model as compared to the Weibull ACD model.

In `R`, the *ACDm* package, created by Belfrage (2016), can be used to fit autoregressive conditional duration models and several extensions thereof, including the Log ACD model, additive and multiplicative ACD model (Hautsch, 2011), augmented box-Cox model (Hautsch, 2011) and spline news impact ACD model (Hautsch, 2011). The function *diurnalAdj* is used to create diurnally adjusted duration series, while the function *acdFit* allows different specifications for the error distribution and fits the models primarily using maximum likelihood estimation.

5 Practical Challenges in HFT Applications

HFT data pose a set of new challenges for researchers and practitioners as mentioned in Section 1. In this section, we discuss a few interesting

applications of HFT data and mention a few open challenges that are yet to be resolved. Comprehensive statistical modeling and inference of the duration process can give insight into the market's buyer and seller trading activity patterns, which is a topic of considerable interest in market microstructure theory. One of the important questions in this context is how to best construct and model inter-event durations. Long range dependence (persistent memory) in durations is also an important stylized feature of HFT data, and adequate modeling of persistence is still an open problem. Furthermore, the asynchronicity problem poses several challenges in multivariate modeling of HFT data, including covariance estimation of multiple assets, which has several applications in trading, risk management and portfolio rebalancing. The following sections discuss these issues.

5.1. Constructing and Modeling Durations High-frequency trading has yielded massive amounts of dense irregular data, leading to the question of how best to subsample or aggregate the data to facilitate effective data analysis. Indeed most statistical approaches involve some form of subsampling or aggregation. This not only reduces the amount of data to be processed, but also limits the effect of noise and allows researchers to create economically meaningful variables (e.g., realized volatility). Many practitioners subsample the data according to clock time; for example, Bloomberg allows users to retrieve intra-day data over fixed time intervals ranging from 1 min to 24 h. For each interval, referred to as a bar, there are several associated quantities, including the volume traded over the bar, and the opening and closing prices. These values can then be used to compute variables of interest; for instance, closing prices over successive bars can be used to calculate squared returns, which are then averaged to produce an estimate of volatility.

Time bars, however, can have undesirable empirical and theoretical properties. For instance, when data is sampled uniformly in time, the resulting price changes often display volatility clustering (i.e., large price changes tend to be followed by other large price changes), which makes modeling more challenging. Furthermore, HF traders may have little need for data analysis that is performed with time bars: forecasts over fixed time horizons (e.g., 1 h) are less useful than forecasts over fixed volumes (e.g., 100,000 shares), since execution algorithms typically consider volume without regard to how quickly or slowly that volume can be traded (Easley et al., 2012a).[1]

[1]Indeed this is one of the criticisms of HFT. The May 6, 2010 "Flash Crash," in which the Dow Jones Industrial Average dropped by almost 1,000 points in 30 min, was the result of an execution algorithm that considered only volume, not time. As a result, $4.1 billion of E-Mini S&P 500 futures contracts were sold on the Chicago Mercantile Exchange in a mere 20 min interval (Goldstein et al. (2014)).

Event aggregation, i.e., aggregation of a process based on some specific trading event, is one of the most popular approaches to forming durations. There are many ways of defining events and hence inter-event durations, e.g., price durations, volume durations, trade durations etc. (Hautsch, 2011). In constructing trade durations for stocks with low liquidity, we may define events as successive trades. For more liquid stocks, however, we need to use alternative event definitions to reduce the impact of market microstructure effects. In what follows, we discuss two such approaches for constructing durations: the price threshold approach and the dollar-volume approach. The two methods differ in how events are defined. The price threshold approach considers the price at which trades are executed, while the dollar-volume approach incorporates information both on price and volume traded.

Defining Events Based on Price Threshold Under this method, an event is said to have occurred when the change in price between successive trades exceeds a certain threshold, say δ_p. Consider a particular trading day. Let P_0 be the opening price of the stock on that day and let P_i denote the price of the stock at time t_i. Let the total number of trades that day be denoted by T. Then we define the set of events as

$$E = \{i : |P_i - P_{i-1}| > \delta_p,\ i = 1, \ldots T\}. \tag{5.1}$$

Typically modelers choose the threshold, δ_p, in such a way that it captures realistic price movements. For instance, Bauwens and Giot (2000) use a value of \$0.125 and Engle and Russell (1998) set δ_p to be one-half the largest observed price spread (see Fig. 2, left).

Defining Events Using Dollar-Volume Bars The dollar-volume method defines an event as occurring when the change in cumulative dollar-volume exceeds a given threshold, δ_v. Let P_j and V_j denote the price and volume of the stock at trade j. Their product, P_jV_j, is then the dollar-volume of trade j. If event $(i-1)$ occurs with the execution of trade r_{i-1}, then event i takes place whenever trade r_i occurs, where

$$r_i := \arg\min_{r^*} \left\{ \sum_{j=r_{i-1}+1}^{r^*} P_jV_j \,\middle|\, \sum_{j=r_{i-1}+1}^{r^*} P_jV_j \geq \delta_v \right\}. \tag{5.2}$$

In other words, r_i is the first trade at which the total dollar-volume (summed over all trades that took place since event $i-1$) exceeds the threshold, δ_v

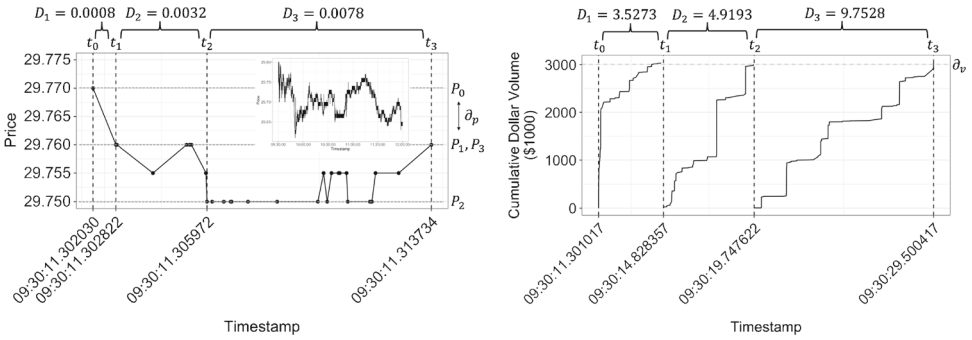

Figure 2: Constructing durations for Bank of America (BAC) based on trades that took place around the market opening on January 2, 2018. **Left**: Under the price threshold approach, an event occurs when the price change exceeds some threshold, δ_p. Trades are marked with circles and the first three inter-event durations (D_1 -D_3) are shown. A plot of BAC's full price trajectory is inlaid. **Right**: Under the dollar volume approach, an event occurs when the cumulative dollar volume exceeds some threshold, δ_v. As on the left, the first three inter-event durations are marked

(see Fig. 2, right). The i^{th} dollar-volume bar consists of all trades between events $(i-1)$ and i. Notice that δ_v varies from stock to stock, with larger firms tending to have greater daily dollar-volume and thus higher thresholds. Moreover, a stock will have fewer events on days when it is relatively inactive and more events on days when it is highly traded. The selection of the threshold δ_v to be $\frac{1}{100}$ would mean that on average each stock will have 100 dollar-volume bars each day.

The dollar-volume approach for constructing durations has several appealing properties. First, Easley et al. (2012a, b) show that this technique yields series of price changes that exhibit less heteroscedasticity and less serial correlation, and whose distribution is closer to normal, than what results from sampling in clock time. With approximate normality and independence, practitioners can use standard statistical methods, which may be faster to implement and thus provide a time advantage over competitors (Easley et al., 2012a). Second, Easley et al. (2021) argue that sampling by volume acts as a proxy for sampling by information content. A trade's dollar-volume is correlated with the amount of news entering the market; thus, by defining events based on a constant amount of cumulative dollar-volume, we ensure that each bar represents the same amount of underlying information. Dollar-volume sampling has been used to address a variety of

market microstructure questions, e.g., to improve on PIN (Probability of Informed Trading) (see Section 6).

Empirical Analysis of Durations We demonstrate two approaches for constructing durations (price threshold and dollar-volume threshold) on real data. We use HF transaction-by-transaction stock price data for two assets, Bank of America (BAC) and 3M (MMM). This data was obtained from the Trade and Quotes (TAQ) database at Wharton Research Data Services (WRDS) (N.Y.S.E. Trade and Quote Database, 2019). We consider data from all trading days in June 2018 ($N = 21$). For each of these days, raw durations were obtained from the transactions data as $D_i = t_i - t_{i-1}$, where we have assumed the occurrence of an event occurs in two different ways, as described below. BAC is a highly liquid stock with trading volumes (100,000 to 200,000 transactions per day on average) and with a relatively low price spread of about \$0.5 per day on average. On the other hand, MMM is a relatively lesser liquid stock with trading volumes (15,000 to 30,000 transactions per day on average) and with a price spread of about \$3 per day on average.

- Price threshold: as in the method described earlier, we compute the threshold using data from January 2018 to May 2018. For each month, we calculate the average turnover ratio as the average daily volume, divided by total shares outstanding. Then we compute the average of those values to obtain δ_p (see Eq. 5.1). We find that δ_p is 0.0054 for BAC and 0.0040 for MMM. As expected, the lengths of the duration time series depend on the liquidity of the stock.

- Dollar-volume threshold: using the data from January 2018 to May 2018, we calculate the mean daily dollar-volume. Then, we set the value of δ_v (see Eq. 5.2) to be $\frac{1}{700}$th of the stock's mean daily dollar-volume. Note that we do not use a fraction of $\frac{1}{50}$th, as originally suggested in Easley et al. (2012a). This is because we experience convergence issues when we try to fit multiplicative error models to a time series with less than 500 data points.

Using price and dollar-volume thresholds, we obtain for each stock two time series of different lengths for each trading day. For example, for BAC on June 6, 2018, we obtain a time series of length 13,181 using the price threshold approach and a time series of length 696 using the dollar-volume threshold approach. We fit different ACD(p, q) models to the adjusted durations, select the best model based on the Bayesian Information Criterion

(BIC), and use the best fitted model to calculate in-sample errors. More specifically, we have fitted 16 different ACD(p, q) models using Weibull innovations, where p, q can take values in $\{0, 1, 2, 3, 4\}$. We fit these models for each trading day in June 2018 and for both stocks.

Table 1 shows the in-sample errors for BAC and MMM, using different threshold methods. The in-sample error is calculated as the relative mean absolute difference between the actual adjusted durations and the fitted adjusted durations. In the case of MMM, in-sample performance using the price threshold method is superior, while in the case of BAC, we observe that the dollar-volume and price threshold methods exhibit comparable performance.

5.2. Persistence in Durations Long memory behavior in time series has been well documented in Beran (1994), Palma (2007), & Robinson (2003). Although long memory patterns have been explored extensively in return and volatility series (Baillie, 1996), only a few approaches have focused on such effects in HFT data. Long range dependence is a key stylized feature in time series of durations as noted in Section 3. Ever since the seminal paper of Engle and Russell (1998), who proposed ACD models for modeling durations and also provided empirical evidence for the persistence in durations, a number of studies have documented the slowly decaying autocorrelation function of transaction, price, and volume durations. More recently, studies of durations in Sun et al. (2008), Deo et al. (2010), Chen et al. (2013), Cartea and Jaimungal (2013), and Žikeš et al. (2017) have produced compelling evidence of long memory in equities and currencies.

Jasiak (1999) proposed a class of fractionally integrated ACD (FIACD) models, (which are analogous to fractionally integrated GARCH (FIGARCH) models proposed by Baillie et al. (1996)), and studied long memory behavior in IBM trade durations. Hautsch et al. (2006) generalized the idea of long memory in the Log ACD model, referred to as Long memory Log ACD (LM-Log ACD) model. The LM-Log ACD(p, d, q) model is specified as

$$\begin{aligned} x_i &= e^{\psi_i}\epsilon_i, \\ \psi_i &= \omega + (1 - \beta(L))^{-1}(1 - L)^{-d}\alpha(L)\epsilon_i, \end{aligned} \tag{5.3}$$

where $\beta(L) = \sum_{i=1}^{q} \beta_i L^i$, $\alpha(L) = \sum_{i=1}^{p} \alpha_i L^i$ denote polynomials of lag operator L and ϵ_i follows an i.i.d. process defined on the positive support with $E[\epsilon_i] = 1$. The parameter d is the long memory parameter with $d \in (-0.5, 0.5)$. The stationarity conditions for the LM-Log ACD are derived in Feng and Zhou (2015). The Log ACD$_2(1, 1)$ model in Eq. 4.21 is obtained by letting $d = 0$, $p = 1$ and $q = 1$ in Eq. 5.3.

Table 1: In-sample errors for BAC and MMM, using both dollar-volume and price threshold durations

Day of week	Date	BAC		MMM	
		Dollar-Volume Threshold	Price Threshold	Dollar-Volume Threshold	Price Threshold
Friday	2018-06-01	0.51%	0.97%	2.10%	0.99%
Monday	2018-06-04	0.98%	0.94%	1.07%	0.34%
Tuesday	2018-06-05	0.98%	0.90%	1.47%	1.08%
Wednesday	2018-06-06	0.06%	1.21%	1.02%	0.51%
Thursday	2018-06-07	1.05%	1.12%	2.84%	0.80%
Friday	2018-06-08	1.55%	0.91%	0.38%	0.91%
Monday	2018-06-11	1.44%	1.03%	1.57%	0.83%
Tuesday	2018-06-12	1.53%	0.88%	0.27%	0.72%
Wednesday	2018-06-13	0.05%	1.25%	1.60%	0.26%
Thursday	2018-06-14	1.24%	1.00%	0.79%	1.07%
Friday	2018-06-15	1.78%	0.91%	2.60%	0.50%
Monday	2018-06-18	2.26%	1.00%	0.86%	0.51%
Tuesday	2018-06-19	0.70%	1.13%	3.13%	0.66%
Wednesday	2018-06-20	1.12%	1.02%	1.16%	0.71%
Thursday	2018-06-21	0.18%	1.14%	1.71%	0.60%
Friday	2018-06-22	0.50%	1.03%	1.73%	0.88%
Monday	2018-06-25	2.41%	0.91%	0.86%	1.22%
Tuesday	2018-06-26	0.24%	1.16%	0.38%	0.97%
Wednesday	2018-06-27	0.91%	0.97%	1.64%	0.43%
Thursday	2018-06-28	1.21%	0.76%	1.39%	0.90%
Friday	2018-06-29	0.31%	1.82%	0.53%	1.05%

We present preliminary results from an investigation of persistence in durations using the same data that we used in Section 5.1, i.e., dollar-volume threshold and price threshold based durations for BAC and MMM. We fit the LM-Log ACD $(1, d, 0)$ model in Eq. 5.3 to the durations data using our R code (available upon request) and obtain the conditional MLEs of the unknown model parameters including the persistence parameter d. Table 2 shows the estimated values of d with their estimated standard errors. These results show that the estimated persistence for BAC (higher liquidity) is higher than that of MMM (lower liquidity) for any day, both for the dollar-volume threshold and price threshold based durations. A day-of-the-week

Table 2: Estimated values of d from LM-Log ACD$(1, d, 0)$ model and their standard errors in brackets for two types of threshold. $\hat{d}_{DV}$ is the estimate of d for Dollar-Volume threshold whereas $\hat{d}_P$ is the estimate of d for Price threshold

Day of	Date	BAC		MMM	
Week		$\hat{d}_{DV}$	$\hat{d}_P$	$\hat{d}_{DV}$	$\hat{d}_P$
Friday	2018-06-01	0.241 (0.038)	0.226 (0.016)	0.117 (0.052)	0.189 (0.027)
Monday	2018-06-04	0.225 (0.055)	0.210 (0.024)	0.172 (0.037)	0.133 (0.03)
Tuesday	2018-06-05	0.159 (0.053)	0.262 (0.027)	0.108 (0.054)	0.187 (0.029)
Wednesday	2018-06-06	0.262 (0.048)	0.172 (0.021)	0.179 (0.041)	0.224 (0.031)
Thursday	2018-06-07	0.242 (0.038)	0.234 (0.013)	0.102 (0.057)	0.162 (0.041)
Friday	2018-06-08	0.154 (0.056)	0.215 (0.021)	0.128 (0.044)	0.128 (0.038)
Monday	2018-06-11	0.218 (0.062)	0.221 (0.024)	0.172 (0.036)	0.116 (0.053)
Tuesday	2018-06-12	0.145 (0.076)	0.105 (0.037)	0.156 (0.049)	0.085 (0.055)
Wednesday	2018-06-13	0.201 (0.035)	0.154 (0.017)	0.147 (0.054)	0.053 (0.043)
Thursday	2018-06-14	0.244 (0.027)	0.204 (0.017)	0.175 (0.039)	0.123 (0.033)
Friday	2018-06-15	0.236 (0.034)	0.231 (0.019)	0.242 (0.026)	0.119 (0.026)
Monday	2018-06-18	0.268 (0.038)	0.251 (0.020)	0.137 (0.049)	0.182 (0.026)
Tuesday	2018-06-19	0.266 (0.053)	0.195 (0.016)	0.158 (0.039)	0.218 (0.026)
Wednesday	2018-06-20	0.271 (0.049)	0.228 (0.015)	0.164 (0.042)	0.221 (0.019)
Thursday	2018-06-21	0.138 (0.044)	0.160 (0.024)	0.164 (0.047)	0.184 (0.032)
Friday	2018-06-22	0.192 (0.037)	0.173 (0.016)	0.135 (0.054)	0.161 (0.03)
Monday	2018-06-25	0.170 (0.041)	0.220 (0.023)	0.129 (0.04)	0.077 (0.034)
Tuesday	2018-06-26	0.239 (0.041)	0.244 (0.014)	0.201 (0.035)	0.099 (0.036)
Wednesday	2018-06-27	0.179 (0.038)	0.190 (0.018)	0.098 (0.057)	0.082 (0.043)
Thursday	2018-06-28	0.134 (0.074)	0.199 (0.019)	0.14 (0.063)	0.174 (0.023)
Friday	2018-06-29	0.216 (0.027)	0.204 (0.019)	0.241 (0.028)	0.131 (0.031)

analysis similar to Zhang et al. (2019) shows that that on average, the persistence appears to be higher (lower) for BAC at the start (end) of the week, while the reverse is true for MMM. More detailed analysis following these preliminary results for durations will be a useful avenue of research.

Other avenues of research can also be pursued along the following directions. Deo et al. (2010) proposed a long memory version of stochastic conditional duration (SCD) model (Bauwens and Veredas, 2004) known as Long-memory Stochastic Duration model (LMSD) by letting the latent factor to follow a long-memory process. Empirical analyses show that LMSD is a better fit than the autoregressive conditional duration (ACD) model. Thavaneswaran et al. (2015) considered modeling long range dependence in durations using long memory stochastic conditional duration (LMSCD) model, which is defined along the lines of FIACD model. There is an emerging body of literature which considers modeling persistence in pure jump processes (Cao et al., 2017; Hsieh et al., 2019).

5.3. Multivariate Relationship with Asynchronous Data Asynchronicity is one of the stylized features of HFT data as referred to in Section 3.

Tick-by-tick transactions of assets are not homogeneously spaced like low-frequency (e.g., daily) time series. Instead, they usually occur randomly and asynchronously and are accompanied by microstructure noise. As a consequence, it is not straightforward to apply the existing multivariate time series models to intraday data and hence the covariance estimation is challenging. In econometrics literature, the price dynamics of high frequency assets are known to be characterized by 'lead-lag' effects, which means that some assets (laggers) tend to follow the movements of other assets (leaders). This is an important phenomenon as noted in the empirical finance literature (Chan, 1992; De Jong and Nijman, 1997; Dobrev and Schaumburg, 2017) and in the statistics literature (Hoffmann et al., 2013; Hayashi and Koike, 2017) among many others. The estimation of contemporaneous and lagged correlations among assets traded at high-frequency is more complex than with lower frequency (e.g., daily) data, due to asynchronous trading, which prevents the usage of traditional methods. The asynchronous nature of trading results in two main types of spurious lead-lag correlations when standard estimators are used (Buccheri et al., 2021b). First, due to frequent trading activity of some assets, they seem to lead other assets. This effect is due to different levels of trading activity and is not necessarily related to cross-asset pricing. Second, even when the assets are traded at similar levels, there exist spurious nonzero lead-lag correlations that are unrelated to true lead-lag dependencies. A combination of autocorrelation and contemporaneous correlations might also be a source of spurious lead-lag correlations.

Estimating Contemporaneous Relationships Among Multiple Stocks from HFT Data In high-frequency finance, the problem of estimating and forecasting intraday volatilities and correlations is of prime importance. For example, a high-frequency trader may be interested in rebalancing the portfolio on an intraday basis and therefore requires accurate forecasts of short-term covariance. Likewise, presence of highly correlated stocks as constituents in a portfolio might increase the probability of a large loss and hence precise estimation of the covariance matrix of the stocks is necessary. Additionally, the study of intraday dependencies of financial assets offers insight into the market's reaction to external information and is theoretically relevant to the study of market microstructure.

Several conditional covariance models have been proposed in econometric literature for regularly spaced time series, and they are widely used in risk and portfolio management at daily or lower frequencies. Popular multivariate dynamic time-series models include the class of multivariate

extensions of the univariate GARCH model of Engle (1982) and Bollerslev (1986), the Dynamic Conditional Correlation (DCC) model of Engle (2002a) and multivariate stochastic volatility model (MSV) of Harvey et al. (1994). A limitation of these models is that they are misspecified in cases where data are recorded with observational noise and require synchronization when data are irregularly spaced. Therefore, it is not straightforward to apply these models to HFT data, since the HF prices are contaminated by microstructure noise and assets are traded asynchronously. These effects may lead to significant data loss and underestimation of correlations (Buccheri et al., 2021a).

Often analysts have trouble dealing with the multiple HFT data at once when conducting multivariate analysis since assets do not trade on a fixed grid, trades and quotes don't arrive synchronously. Hence synchronization of HF data from multiple assets is necessary. Data synchronization includes explicit schemes, such as previous tick, refresh time (Barndorff-Nielsen et al., 2011), generalization sampling time (Aït-sahalia et al., 2010), implicit approaches used in Hayashi-Yoshida estimator Hayashi et al. (2005), and preaveraging estimators (Christensen et al., 2010; Jacod et al., 2009). Here we give a brief review of the refresh time sampling procedure. More detailed discussion about the other synchronization procedures can be found in (Wang and Zou, 2014). We also illustrate refresh time sampling procedure using three stocks Intel, General Electric (GE) and Cisco (Fig. 3). Assume there are p stocks, and trading time of the ith stock is given by $t_{i\ell}, \ell = 1, \ldots, n_i, i = 1, \ldots, p$. For a given time t, define N_t^i = the number of $t_{ie} \leq t$, $\ell = 1, \cdots, n_i$, which counts the number of distinct data points $t_{i\ell}$ available for the ith asset up to time t. The first refresh time is defined as $\tau_1 = \max\{t_{11}, \ldots, t_{p1}\}$, which is the first time taken to trade all assets

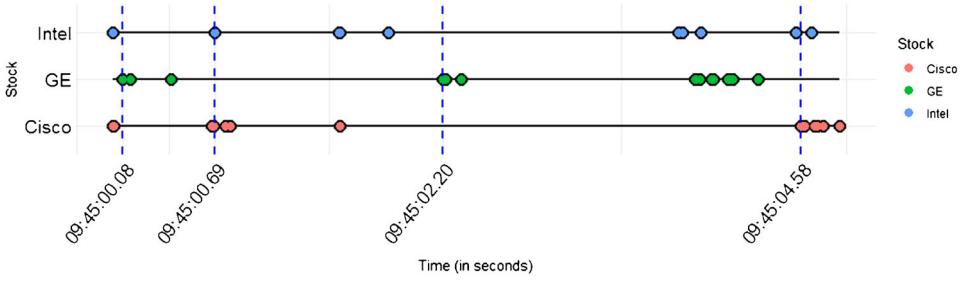

Figure 3: Refresh Time Sampling for Intel, GE and Cisco.The dotted vertical blue lines represent the refresh time points

and refresh their posted prices. The subsequent refresh times are defined as follows. Given the jth refresh time τ_j, define the $(j+1)$th refresh time

$$\tau_{j+1} = \max\left\{t_{1,N^1_{\tau_j}+1}, \ldots, t^p_{p,N^p_{\tau_j}+1}\right\} \tag{5.4}$$

Suppose there are m refresh time points $\tau_1, \cdots, \tau_m$. Intuitively, τ_2 is the second time when all the assets are traded and their prices are refreshed. The following figure illustrates the refresh time sampling idea and in this example $\tau_1 = 09:45:00.08, \tau_2 = 09:45:00.69$, $\tau_1 = 09:45:02.20$ and $\tau_4 = 09:45:04.58$ are the first four refresh times. The refresh time sampling procedure reduces the Epps effect (Boudt et al., 2021b). A limitation of refresh time sampling procedure is that the least liquid stock determines the sampling grid and one tend to lose many observations and even with same liquidity, because of random arrivals, large data losses may occur in high dimensions.

Although intraday covariance estimation from HFT data is a difficult task due to asynchronicity, there have been several studies which has been devoted to the estimation of the covariance from HFT data (Aït-sahalia et al., 2010; Zhang, 2011; Mancino and Sanfelici, 2011; Bibinger, 2011; Corsi and Audrino, 2012; Peluso et al., 2014; Buccheri et al., 2021a). More recently, there has been an interest in exploring non-linear dependence structures in high frequency asset returns through copula (Chakrabarti and Sen, 2019).

Estimating Lead-Lag Relationships Among Multiple Stocks from HFT Data Understanding lead-lag relationships between the time series of different stocks' returns or volatilities can provide insight into the underlying network structure of an interlinked financial market. For regularly spaced daily or monthly time series, such strategies have received significant interest in the empirical finance and econometrics literature (Billio et al., 2012; Diebold and Yılmaz, 2014). There is robust empirical evidence that networks built on lead-lag relationships tend to be denser during market downturns and systemic events such as a financial crisis, and can be useful in monitoring systemic risk build-up in the financial markets. While building similar networks based on HFT data can potentially provide deeper insight into important linkages in financial markets, the asynchronous nature of HFT data poses considerable challenge in developing statistical techniques to measure such lead-lag relationships. To the best of our knowledge, common methods such as Granger causality (Granger, 1969) have not been generalized to build financial networks from HFT data. Developing rigorous statistical methods for measuring lead-lag relationships from asynchronous HFT data of multiple stocks is a promising research direction.

6 Econometric Perspectives of HFT

Academics and practitioners have used HFT data in a range of applications. Some of these applications involve testing competing hypotheses about market microstructure or assessing the likelihood of informed trading in a financial market. Other applications are more econometric in nature, where HFT data is used to provide intra-day versions of commonly used risk assessment measures such as Value-at-Risk (VaR) and Expected Shortfall (ES).

Testing Theory About Market Microstructure Tay et al. (2004) introduces autoregressive conditional marked durations (ACMD) models to analyze marked duration processes, using events such as tick movements and trade directions (buy/sell) as marks. This model helps explain how trade direction, size, and frequency are transmitted into prices through durations.

Let there be $N+1$ transactions (events) and let t_i denote the time of occurrence of the ith event. Assume that there are m discrete states of the marks associated with $\{t_i\}$, and let the state of the mark when the ith event occurs be denoted by W_i (and its realization w_i), $i = 1, \ldots N$. These marks are useful in classifying each event occurrence, so that there are m underlying stochastic processes, each governing one of these m states of the mark. The flexibility of ACMD models lies in the fact that other exogenous market variables can be augmented into the model, depending on the market hypotheses we want to examine. Denote the vector of exogenous variables, observed after the ith event, occurs by $\boldsymbol{v}_i$, where $i = 1, \ldots N$. The information set after the ith event is $\Phi_i = \{t_r, w_r, \boldsymbol{v}_r; r = 1, 2, \ldots, i\}$. Let the duration between the $(i-1)$th and ith events be X_i (with realizations x_i, $i = 1, \ldots, N$). The adjusted durations are obtained after the data cleansing and diurnal effect adjustments described in Section 3.

Let T_{ji} be the random time duration between the $(i-1)$th and ith events when the state j of the mark is observed, and assume that—conditional on the past information set, Φ_{i-1}—the T_{ji}'s are independent over j, where $j = 1, \ldots m$. The joint distribution of the mark, W_i, and duration, X_i, can be expressed as

$$p_i(k, x_i|\Phi_{i-1}) = P(W_i = k \cap X_i = x_i|\Phi_{i-1}) = \prod_{j\in\Omega} S_{T_{ji}|\Phi_{i-1}}(x_i).\frac{f_{T_{ki}|\Phi_{i-1}}(x_i)}{S_{T_{ki}|\Phi_{i-1}}(x_i)} \tag{6.1}$$

for all $k \in \Omega = \{k_1, \ldots, k_m\}$, where $f_{T_{ki}|\Phi_{i-1}}(x_i)$ and $S_{T_{ki}|\Phi_{i-1}}(x_i)$ are the density function and the survival function of $T_{ji}|\Phi_{i-1}$, respectively. With

appropriate distributional assumptions on $T_{ji}|\Phi_{i-1}$, one can derive the explicit form of the joint distribution and marginals of X_i and W_i, conditional on the past information set, Φ_{i-1}. The log-likelihood function for the ACMD model is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p_i(k, x_i|\Phi_{i-1}) = \sum_{i=1}^{N} \left[\sum_{j\in\Omega} \ln S_{T_{ji}|\Phi_{i-1}}(x_i) + \ln \frac{f_{T_{w_i,i}|\Phi_{i-1}}(x_i)}{S_{T_{w_i,i}|\Phi_{i-1}}(x_i)} \right],$$

where $\boldsymbol{\theta}$ denotes the set of parameters of the underlying distribution. Let $E(T_{ji}|\Phi_{i-1}) = \psi_{ji}$ be the expected marked duration, which is governed by m ACD models, as discussed in Section 4.4. Then ψ_{ji} is updated according to the following dynamics:

$$\ln \psi_{ji} = \sum_{k\in\Omega} v_{jk} D_k(w_{i-1}) + \alpha_j \ln \psi_{j,i-1} + \beta_j \ln x_{i-1} + f(x_{i-1}, \boldsymbol{v}_{i-1}; \boldsymbol{\rho}_j) \quad (6.2)$$

for all $j \in \Omega$. The function f is linear in the natural logarithm of the exogenous variables in $\boldsymbol{v}$, contains possible interactions with x_{i-1}, and each term is weighted by elements in the vector $\boldsymbol{\rho}_j$ The variable $D_k(z)$ is an indicator variable of the state k for the mark, which assumes value 1 if $z = k$ and 0 otherwise, $k \in \Omega$. The parameters of this model can be estimated using the maximum likelihood approach.

Tay et al. (2004) uses ACMD models to study competing hypotheses from Easley et al. (2002) and Diamond and Verrecchia (1987) about whether low transaction rates indicate bad news or no news. The empirical results are in favor of Diamond and Verrecchia (1987). This model has been used for estimating the probability of informed trading (PIN), as described in the next subsection, using high-frequency transaction data. Kwok et al. (2009) demonstrates an application of ACMD models using stocks traded on the Hong Kong Stock Exchange (SEHK).

Estimating the Probability of Informed Trading (PIN) PIN is defined as the probability that a counterparty in the trading process has private information on the value of the asset exchanged. It serves as a proxy for the proportion of informed traders in the market. The parameters needed to compute PIN are obtained from the estimation of a theoretical model of the trading process. This is a key concept in market microstructure theory and is widely used as an indicator of asymmetric risk information related to asset trading. Easley et al. (1996, 2002) build a structural model for the trading process which assumes that the market is composed of a heterogeneous population of traders, with informed traders, market makers, and uninformed

traders. For a given stock on the ith trading day, let B_i denote the buy orders and S_i denote the sell orders. An information event for a given stock is assumed to follow a Bernoulli distribution with success probability α. This event reveals either a signal for a low stock value (with probability δ) or a signal for a high stock value (with probability 1 - δ). The informed traders enter the market when an information-revealing event happens and they are assumed to place buy (sell) orders at a rate of μ. Uninformed traders are assumed to place buy orders at a rate of ϵ_b and sell orders at a rate of ϵ_s, independent of the information event and the signal. The orders from the informed and uninformed traders are both assumed to follow Poisson processes.

For the ith trading day, the joint probability distribution of (B_i, S_i), given the parameter vector $\Theta = (\alpha, \delta, \mu, \epsilon_b, \epsilon_s)$, is

$$\begin{aligned} f(B_i, S_i|\Theta) &= \alpha\delta \exp(-\epsilon_b)\frac{\epsilon_b^{B_t}}{B_t!}\exp(-(\epsilon_s+\mu))\frac{(\epsilon_s+\mu)^{S_t}}{S_t!} \\ &\quad +\alpha(1-\delta)\exp(-(\epsilon_b+\mu))\frac{(\epsilon_b+\mu)^{B_t}}{B_t!}\exp(-\epsilon_s)\frac{\epsilon_s^{S_t}}{S_t!} \\ &\quad +(1-\delta)\exp(-\epsilon_b)\frac{\epsilon_b^{B_t}}{B_t!}\exp(-\epsilon_s)\frac{\epsilon_s^{S_t}}{S_t!}. \end{aligned} \tag{6.3}$$

Hence, by assuming the trading activity to be independent across T days, the log-likelihood is given by

$$L(\Theta|T) = \sum_{i=1}^{T} \log(f(B_i, S_i|\Theta)).$$

Using the maximum likelihood approach, along with the boundary conditions $\alpha, \delta \in [0, 1]$ and $\mu, \epsilon_b, \epsilon_s \in [0, \infty)$, we obtain the estimates $\hat{\Theta} = (\hat{\alpha}, \hat{\delta}, \hat{\mu}, \hat{\epsilon_b}, \hat{\epsilon_s})$. The PIN estimate is then given by

$$\widehat{\text{PIN}} = \frac{\hat{\alpha}\hat{\mu}}{\hat{\alpha}\hat{\mu} + \hat{\epsilon_b} + \hat{\epsilon_s}}. \tag{6.4}$$

Thus PIN is the ratio of the expected number of trades per day initiated by informed traders to the expected total number of trades per day.

Dollar-volume bars can be used to calculate an extension of PIN known as the volume-synchronized probability of informed trading (VPIN), introduced in Easley et al. (2012b). By comparing the amount of buyer- and seller-initiated trades,[2] VPIN measures the extent to which there is information

[2]A number of approaches can be used to classify trades as buyer- or seller-initiated, including the Lee-Ready algorithm, the tick rule, and bulk volume classification (see Easley et al. 2016 and references therein).

asymmetry in the market. (For example, if a group of traders knows that a stock's price is about to rise, we may observe a preponderance of buyer-initiated trades.) The VPIN at bar i is given by

$$\text{VPIN}_i = \frac{1}{W} \sum_{k=i-W+1}^{i} \frac{|\hat{V}_k^S - \hat{V}_k^B|}{V_k}, \tag{6.5}$$

where V_k is the total volume traded over bar k, $\hat{V}_k^B$ is the estimated total buyer-initiated volume over bar k, $\hat{V}_k^S = V_k - \hat{V}_k^B$ is the estimated total seller-initiated volume over bar k, and W is the length of a lookback window. Notice that VPIN can be calculated directly from trade data, whereas computing PIN requires us to first estimate unobservable parameters of a theoretical microstructure model: α, μ, ϵ_b, and ϵ_s in Eq. 6.3.

Estimating Intra-day Value-at-Risk (IVaR) Value at Risk (VaR) is a popular risk management tool used by financial institutions to quantify the level of financial risk within a firm, portfolio, or position, over a specific time frame. It measures the worst expected loss of a risky asset over a certain period of time and at a given confidence level. Formally, VaR can be defined as the conditional quantile of the asset return distribution for a given horizon and a given shortfall probability, α, whose value is typically between 1% and 5%.

Let r_t denote the return of the asset over the time period $t-1$ to t. The ex-ante VaR forecast with a target probability of α solves the following:

$$P_{t-1}^M(r_t < -\text{VaR}_t(\alpha)) = \alpha, \tag{6.6}$$

where P_{t-1}^M is the probability derived from model M using the information up to time $t-1$. The negative sign in the above equation is due to the convention of reporting VaR as a positive number. Unlike the daily VaR, the use of high-frequency data in computing intra-day VAR (IVaR) poses many challenges, including irregular spacing and intra-day periodicity. Details of the stylized facts of high-frequency data has been highlighted in Section 3 of this paper.

Giot (2005) considers the estimation of IVaR using equally spaced intra-day returns, employing Normal GARCH, t-GARCH, and RiskMetrics models for comparison. Using irregularly spaced tick-by-tick data, Dionne et al. (2009) proposes a Monte Carlo simulation procedure to estimate IVaR. Coroneo and Veredas (2012) proposes quantile regression for regularly spaced high-frequency data. The approaches in Giot (2005) and Dionne et al. (2009)

have some important shortcomings. Giot's method of calculating IVaR is based on regularly spaced time series of returns and hence does not account for the effect of durations. On the other hand, Dionne's method is based on irregularly spaced time series of returns and takes durations into account, but the returns and durations are modeled separately. This imposes restrictions on the behavioral assumption of traders. To overcome these limitations, Liu and Tse (2015) proposes a simulation-based approach to estimate IVaR, which assumes price movement and durations follow a two-state asymmetric autoregressive conditional duration (AACD) model. In this framework, the price movements and durations are modeled jointly. The AACD model has been used to model stock price dynamics in Tay et al. (2011).

Let $\{t_i\}_{i=0}^{N}$ be a sequence of times in which t_i is the time of occurrence of the ith event, which occurs whenever the cumulative change in the logarithmic transaction price exceeds a threshold δ, similar to the price threshold discussed in Section 5.1. Let y_i denote the direction of the price movement of the ith event, such that $y_i \in \{-1, 1\}$, representing downward and upward price movement, respectively. Let $x_{j,i}$, $j = -1, 1$, be the two latent variables corresponding to the two possible states of $y_i = -1$ or $y_i = 1$, respectively. With two possible states at the ith event, there can only be one realized state, which is the shortest of the two latent durations. Let x_i be the observed duration, where $x_i = \min(x_{-1,i}, x_{1,i})$. Let $\psi_{j,i} = E(x_{j,i}|\Phi_{i-1})$, $j = -1, 1$, be the conditional expected duration of the latent variable $x_{j,i}$, with Φ_{i-1} being the information set up to time t_{i-1}. The basic two state AACD model specification is

$$\begin{aligned} x_{j,i} &= \psi_{j,i}\epsilon_{j,i}, \\ \log(\psi_{j,i}) &= \sum_{k=-1,1} (v_{j,k} + \alpha_{j,k}\log(x_{i-1}))D_k(y_{i-1}) + \beta_j \log(\psi_{j,i-1}), \end{aligned} \tag{6.7}$$

where $j = -1, 1$; $i = 1, \ldots, N$; and $D_k(z) = 1$ if $z = k$ and 0 otherwise. Assume that the innovations corresponding to each state ($\epsilon_{-1,i}$ and $\epsilon_{1,i}$ for $i = 1, \ldots, N$) are independently distributed with a Weibull distribution having unit mean. The parameters of the AACD model can be estimated using maximum likelihood estimation.

Suppose we have the estimated model and we want to compute, at time T_1, the IVaR at time T_2 ($T_2 > T_1$), using the simulation-based algorithm presented in Liu and Tse (2015). We can do so by following these steps:

> **Step 0:** Initialize $x_0, \psi_{-1,0}, \psi_{1,0}$, and y_0 based on the information prior to T_1. Then begin the simulation.

Step 1: Set $i = 1, t_0 = 0$ and compute $\psi_{-1,1}$, $\psi_{1,1}$ using Eq. 6.7.

Step 2: Randomly draw $\epsilon_{-1,i}$ and $\epsilon_{1,i}$ from independent Weibull distributions with shape parameters $\widehat{\phi_{-1}}$ and $\widehat{\phi_1}$, respectively. Compute $x_{j,i} = \psi_{j,i}\epsilon_{j,i}$ and $\psi_{j,i+1}$ for $j = -1, 1$. Set $x_i = \min\{x_{-1,i}, x_{1,i}\}$ and $y_i = j$, corresponding to the shorter $x_{j,i}$, where $j = -1, 1$.

Step 3: Collect the time, $t_i = t_{i-1} + x_i$, and set $\log p_i = \log p_{i-1} + j\delta$ for the observed j value in the previous step. Note that δ is the threshold mentioned earlier in this subsection.

Step 4: Set $i = i + 1$ and iterate the second and the third steps until obtaining the first t_i that exceeds $T_2 - T_1$. At this point we obtain a simulated return over the interval (T_1, T_2).

These simulation steps are repeated to obtain an empirical distribution of returns over the interval (T_1, T_2). IVaR(ξ) is obtained by computing the ξ-quantile of the empirical return distribution over the interval (T_1, T_2).

Expected Shortfall Using Intra-day Range Expected Shortfall (ES) is a popular market risk measure that conveys information about the possible exceedances beyond the VaR. Formally, ES can be defined for any loss distribution as

$$\text{ES}(\alpha) = \frac{\int_0^\alpha \text{VaR}(x)dx}{\alpha}, \tag{6.8}$$

which is the average of VaR(x) over all x that are less than or equal to α. Both VaR and ES are popular and important market risk measures, but, despite being widely used, VaR has some significant shortcomings. Firstly, it does not take into account the magnitude of the potential losses beyond VaR. Secondly, it lacks the sub-additive property, i.e., the risk measure for a portfolio can be less than the sum of the risk measures of the components of the portfolio (Gordy and Juneja, 2010). Estimating these risk measures forms an integral part of portfolio optimization (Huang et al., 2010).

One of the fundamental problems of ES estimation is that it is not elicitable, meaning there do not exist scoring functions for its estimation and evaluation. Fissler and Ziegel (2016) shows that VaR and ES are jointly

elicitable and presents a set of consistent joint scoring functions for these two risk measures:

$$h_{FZ}(y_t, q_t, e_t; \theta, G_1, G_2, a) = (\mathbb{1}\{y_t \leq q_t\} - \theta)\left(G_1(q_t) - G_1(y_t) + \frac{1}{\theta}G_2(e_t)q_t\right) - G_2(e_t)\left(\frac{1}{\theta}\mathbb{1}\{y_t \leq q_t\}y_t - e_t\right) - \mathcal{G}_2(e_t) + a(y_t), \tag{6.9}$$

where y_t is the daily return, θ is the probability level, and q_t and e_t are the VaR and ES at the same probability level, θ. Moreover, $G_1, G_2, \mathcal{G}_2$, and a are real-valued functions, G_1 is weakly increasing, G_2 is strictly increasing and positive, and $\mathcal{G}_2' = G_2$. These are referred to as FZ scores, which allow the joint estimation of VaR and ES.

Most earlier methods for forecasting ES did not involve any intra-day information. Recently, though, Gerlach and Chen (2015) and Meng and Taylor (2020) incorporate such information in the form of intra-day range, the difference between the highest and lowest intra-day log prices. One desirable characteristic of the intra-day range is that it is a more efficient volatility estimator than the daily returns (Parkinson, 1980).

Chen (2012) considers several non-linear threshold conditional autoregressive VaR (CAViaR) models that incorporate intra-day price ranges but do not consider ES forecasting. Meng and Taylor (2020) rectifies that by proposing a CAViaR-FZ-Range model for ES predictions, as described by the following equations:

$$\begin{aligned} q_t(\boldsymbol{\beta}) &= \beta_1 + \beta_2 q_{t-1}(\boldsymbol{\beta}) + \beta_3 \text{Range}_{t-1}, \\ e_t(\boldsymbol{\beta}) &= \beta_4 q_t(\boldsymbol{\beta}), \end{aligned} \tag{6.10}$$

where $\boldsymbol{\beta}$ is a vector of parameters and q_t and e_t represent the VaR and ES at the same probability level. This model is estimated by minimizing FZ scores, as in Eq. 6.9.

7 Discussion and Summary

Tick-by-tick, high-frequency data sets have become the new normal in modern financial markets dominated by algorithmic trading. While these data sets provide high-resolution information about the trading process, their analysis poses unique challenges due to stylized features such as irregular spacing. In this paper, we discussed examples of trading behaviors that give rise to such irregular spacing and reviewed existing approaches to model irregularly spaced HFT data. We focused primarily on models for inter-event durations. In addition to reviewing their mathematical expositions

and software implementations, we discussed two methods of defining *events*, one based only on asset price and the other based on both asset price and order volume. We illustrated duration models based on these two types of events using real HFT data sets. We also surveyed some applications of HFT data sets in economics and finance.

In this work, we mainly focused on irregularly spaced *univariate* HFT time series. There is an emerging body of work on analyzing *asynchronous multivariate* HFT data that we have not discussed here. For instance, non-parametric machine learning methods such as biclustering have been explored in the literature, complementing traditional model-based approaches (Liu et al. 2018, 2021). Another important research direction is synchronizing HFT data on multiple assets to perform meaningful investigation of lead-lag patterns. We expect that the vast body of existing work on univariate HFT models will provide a natural starting point to build informative analytic approaches to tackle asynchronous multivariate HFT data sets.

Acknowledgements. The authors are very grateful to the reviewers and editors for their helpful suggestions for improving the paper.

Funding. This paper was based upon work partially supported by the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. In addition, the work of SB was supported in part by an NSF award (DMS-1812128).

Compliance with Ethical Standards.

Conflict of Interest. The authors declare no conflict of interest.

References

AÏT-SAHALIA, Y. and JACOD, J. (2009). Testing for jumps in a discretely observed process. *The Annals of Statistics* 184–222.

AÏT-SAHALIA, Y., FAN, J. and XIU, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association* **105**, 492, 1504–1517.

AÏT-SAHALIA, Y., JACOD, J. and LI, J. (2012). Testing for jumps in noisy high frequency data. *Journal of Econometrics* **168**, 2, 207–222.

ALIZADEH, S., BRANDT, M.W. and DIEBOLD, F.X. (2002). Range-based estimation of stochastic volatility models. *Journal of Finance* **57**, 3, 1047–1091.

ANDERSEN, T.G. and BOLLERSLEV, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance,* **4**, 2-3, 115–158.

ANDERSEN, T.G. and BOLLERSLEV, T. (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39**, 4, 885–905.

ANDERSEN, T.G., BENZONI, L. and LUND, J. (2002). An empirical investigation of continuous-time equity return models. *The Journal of Finance,* **57**, 3, 1239–1284.

ANDERSEN, T.G., BOLLERSLEV, T. and DIEBOLD, F.X. (2007). Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics* **89**, 4, 701–720.

ARDIA, D., BLUTEAU, K., BOUDT, K., CATANIA, L. and TROTTIER, D.-A. (2019). Markov-switching GARCH models in r: the MSGARCH Package. *Journal of Statistical Software* **91**(4).

ASAI, M., CHANG, C. -L. and MCALEER, M. (2017). Realized stochastic volatility with general asymmetry and long memory. *Journal of Econometrics* **199**, 2, 202–212.

BAILLIE, R.T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics,* **73**, 1, 5–59.

BAILLIE, R.T., BOLLERSLEV, T. and MIKKELSEN, H.O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **74**, 1, 3–30.

BARNDORFF-NIELSEN, O.E. and SHEPHARD, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **64**, 2, 253–280.

BARNDORFF-NIELSEN, O.E. and SHEPHARD, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* **2**, 2, 1–37.

BARNDORFF-NIELSEN, O.E. and SHEPHARD, N. (2005). Variation, jumps market frictions and high frequency data in financial econometrics.

BARNDORFF-NIELSEN, O.E. and SHEPHARD, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* **4**, 1, 1–30.

BARNDORFF-NIELSEN, O.E., HANSEN, P.R., LUNDE, A. and SHEPHARD, N. (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics,* **162**, 2, 149–169.

BAUWENS, L. and GIOT, P. (2000). The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks. *Annales d'Economie et de Statistique,* (60):117–149.

BAUWENS, L. and HAUTSCH, N. (2006). Stochastic conditional intensity processes. *Journal of Financial Econometrics* **4**, 3, 450–493.

BAUWENS, L. and VEREDAS, D. (2004). The stochastic conditional duration model: a latent variable model for the analysis of financial durations. *Journal of Econometrics* **119**, 2, 381–412.

BELFRAGE, M. (2016). ACDM: tools for autoregressive conditional duration models. (R package version 1.0.4).

BERAN, J. (1994). Statistics for long-memory processes. CRC Press.

BIBINGER, M. (2011). Efficient covariance estimation for asynchronous noisy high-frequency data. *Scandinavian Journal of Statistics* **38**, 1, 23–45.

BILLIO, M., GETMANSKY, M., LO, A.W. and PELIZZON, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics* **104**, 3, 535–559.

BJURSELL, J. and GENTLE, J.E. (2012). Identifying jumps in asset prices. In: Handbook of computational finance, pp. 371–399. Springer.

BLACK, F. (1976). Studies of stock market volatility changes. In: Proceedings of the American statistical association business and economic statistics section, pp. 177–181.

BLACK, F. (1986). Noise. *The Journal of Finance* **41**, 3, 528–543.

BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 3, 307–327.

BOUDT, K., CORNELISSEN, J., PAYSEUR, S., KLEEN, O. and SJOERUP, E. (2021a). Highfrequency: tools for highfrequency data analysis. https://CRAN.R-project.org/package=highfrequency. R package version 0.9.0.

BOUDT, K., KLEEN, O. and SJØRUP, E. (2021b). Analyzing intraday financial data in r: the highfrequency package. Available at SSRN 3917548.

BROUSTE, A., FUKASAWA, M., HINO, H., IACUS, S., KAMATANI, K., KOIKE, Y., MASUDA, H., NOMURA, R., OGIHARA, T., SHIMUZU, Y. and ET AL (2014). The yuima project: a computational framework for simulation and inference of stochastic differential equations. *Journal of Statistical Software* **57**, 1–51.

BUCCHERI, G., BORMETTI, G., CORSI, F. and LILLO, F. (2021a). A score-driven conditional correlation model for noisy and asynchronous data: an application to high-frequency covariance dynamics. *Journal of Business & Economic Statistics,* **39**, 4, 920–936.

BUCCHERI, G., CORSI, F. and PELUSO, S. (2021b). High-frequency lead-lag effects and cross-asset linkages: a multi-asset lagged adjustment model. *Journal of Business & Economic Statistics,* **39**, 605–621.

CAMERON, A.C. and TRIVEDI, P.K. (2013). *Regression analysis of count data*. Cambridge University Press, Cambridge.

CAO, W., HURVICH, C. and SOULIER, P. (2017). Drift in transaction-level asset price models. *Journal of Time Series Analysis,* **38**, 5, 769–790.

CARR, P. and WU, L. (2003). The finite moment log stable process and option pricing. *The Journal of Finance* **58**, 2, 753–777.

CARR, P., MADAN, D. and CHANG, E. (1998). The variance gamma process and option pricing. *European Finance Review* **2**, 1, 79–105.

CARR, P., GEMAN, H., MADAN, D.B. and YOR, M. (2002). The fine structure of asset returns: an empirical investigation. *The Journal of Business* **75**, 2, 305–332.

CARTEA, A. and JAIMUNGAL, S. (2013). Modelling asset prices for algorithmic and high-frequency trading. *Applied Mathematical Finance,* **20**, 6, 512–547.

CHAKRABARTI, A. and SEN, R. (2019). Copula estimation for nonsynchronous financial data. arXiv:1904.10182.

CHAN, K. (1992). A further analysis of the lead–lag relationship between the cash market and stock index futures market. *The Review of Financial Studies* **5**, 1, 123–152.

CHEN, C.W.S., GERLACH, R., HWANG, B.B.K and MCALEER, M. (2012). Forecasting Value-at-Risk using nonlinear regression quantiles and the intra-day range. *International Journal of Forecasting* **28**, 3, 557–574.

CHEN, F., DIEBOLD, F.X. and SCHORFHEIDE, F. (2013). A Markov-switching multifractal inter-trade duration model, with application to us equities. *Journal of Econometrics* **177**, 2, 320–342.

CHIB, S., NARDARI, F. and SHEPHARD, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics* **108**, 2, 281–316.

CHIB, S., OMORI, Y. and ASAI, M. (2009). Multivariate stochastic volatility. In: Handbook of financial time series, pp. 365–400. Springer.

CHRISTENSEN, K., KINNEBROCK, S. and PODOLSKIJ, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics* **159**, 1, 116–133.

CHRISTENSEN, K., OOMEN, R.C. and PODOLSKIJ, M. (2014). Fact or friction: jumps at ultra high frequency. *Journal of Financial Economics* **114**, 3, 576–599.

cont, r. (2011). Statistical modeling of high-frequency financial data. *IEEE Signal Processing Magazine,* **28**, 5, 16–25.

cont, r. and tankov, p. (2004). *Financial modeling with jump processes.* Chapman & Hall/CRC, Boca Raton.

coroneo, l. and veredas, d. (2012). A simple two-component model for the distribution of intraday returns. *The European Journal Finance* **18**, 9, 775–797.

corsi, f. and audrino, f. (2012). Realized covariance tick-by-tick in presence of rounded time stamps and general microstructure effects. *J. Financ. Econom.* **10**, 591–616.

cox, d.r. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodol.)* **34**, 187–202.

cox, d.r. and oakes, d. (2018). Analysis of survival data. Chapman and hall/CRC.

daley, d.j. and vere-jones, d. (2003). An introduction to the theory of point processes: volume i: elementary theory and methods. Springer.

de jong, f. and nijman, t. (1997). High frequency analysis of lead-lag relationships between financial markets. *J. Empir. Finance* **4**, 259–277.

deo, r., hsieh, m. and hurvich, c.m. (2010). Long memory in intertrade durations, counts and realized volatility of NYSE stocks. *J. Stat. Plan. Inference* **140**, 3715–3733.

diamond, d.w. and verrecchia, r.e. (1987). Constraints on short-selling and asset price adjustment to private information. *J. Financ. Econ.* **18**, 277–311.

diebold, f.x. and yilmaz, k. (2014). On the network topology of variance decompositions: measuring the connectedness of financial firms. *J. Econ.* **182**, 119–134.

dionne, g., duchesne, p. and pacurar, m. (2009). Intraday value at risk (IVaR) using tick-by-tick data with application to the Toronto Stock Exchange. *J. Empir. Finance* **16**, 777–792.

dobrev, d. and schaumburg, e. (2017). High-frequency cross-market trading: model free measurement and applications. Perspectives.

duffie, d., pan, j. and singleton, k. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* **68**, 1343–1376.

dufour, a. and engle, r.f. (2000). Time and the price impact of a trade. *J. Financ.* **55**, 2467–2498.

easley, d. and o'hara, m. (1992). Time and the process of security price adjustment. *J. Financ.* **47**, 577–605.

easley, d., kiefer, n.m., o'hara, m. and paperman, j.b. (1996). Liquidity, information, and infrequently traded stocks. *J. Financ.* **51**, 1405–1436.

easley, d., hvidkjaer, s. and o'hara, m. (2002). Is information risk a determinant of asset returns? *J. Financ.* **57**, 2185–2221.

easley, d., de prado, m.m.l. and o'hara, m. (2012a). The volume clock: insights into the high-frequency paradigm. *J. Portfolio Manag.* **39**, 19–29.

easley, d., lópez de prado, m.m. and o'hara, m. (2012b). Flow toxicity and liquidity in a high frequency world. *Rev. Financ. Stud.* **25**, 1457–1493.

easley, d., de prado, m.l. and o'hara, m. (2016). Discerning information from trade data. *J. Financ. Econ.* **120**, 269–285.

easley, d., lópez de prado, m., o'hara, m. and zhang, z. (2021). Microstructure in the machine age. *Rev. Financ. Stud.* **34**, 3316–3363.

eberlein, e. and keller, u. (1995). Hyperbolic distributions in finance. Bernoulli 281–299.

efron, b. (1986). Double exponential families and their use in generalized linear regression. *J. Am. Stat. Assoc.* **81**, 709–721.

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (2013). Modelling extremal events: for insurance and finance. Springer Science & Business Media.

ENGLE, R. (2002a). Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econ. Stat.* **20**, 339–350.

ENGLE, R. (2002b). New frontiers for ARCH models. *J. Appl. Econom.* **17**, 425–446.

ENGLE, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econom.: J. Econom. Soc.* **50**, 987–1007.

ENGLE, R.F. and RUSSELL, J.R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–1162.

EPPS, T.W. (1979). Comovements in stock prices in the very short run. *J. Am. Stat. Assoc.* **74**, 291–298.

EVANS, K.P. (2011). Intraday jumps and us macroeconomic news announcements. *J. Bank. Finance* **35**, 2511–2527.

FAN, J., LI, Y. and YU, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *J. Am. Stat. Assoc.* **107**, 412–428.

FENG, Y. and ZHOU, C. (2015). Forecasting financial market activity using a semiparametric fractionally integrated log-acd. *Int. J. Forecast.* **31**, 349–363.

FERNANDES, M. and GRAMMIG, J. (2006). A family of autoregressive conditional duration models. *J. Econ.* **130**, 1–23.

FISSLER, T. and ZIEGEL, J.F. (2016). Higher order elicitability and Osband's principle. *Ann. Stat.* **44**, 1680–1707.

FLEMING, T.R. and HARRINGTON, D.P. (2011). *Counting processes and survival analysis*. Wiley, New York.

GERLACH, R. and CHEN, C.W. (2015). Bayesian expected shortfall forecasting incorporating the intraday range. *J. Financ. Econom.* **14**, 128–158.

GERLACH, R. and WANG, C. (2016). Forecasting risk via realized GARCH, incorporating the realized range. *Quant. Finance* **16**, 501–511.

GHALANOS, A. (2020). Rugarch: univariate GARCH models. R package version 1.4-4.

GIOT, P. (2005). Market risk models for intraday data. *Eur. J. Finance* **11**, 309–324.

GOLDSTEIN, M.A., KUMAR, P. and GRAVES, F.C. (2014). Computerized and high-frequency trading. *Financ. Rev.* **49**, 177–202.

GORDY, M.B. and JUNEJA, S. (2010). Nested simulation in portfolio risk measurement. *Manag. Sci.* **56**, 1833–1848.

GRANGER, C.W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econom. J. Econom. Soc.*, 424–438.

HANSEN, P.R. and LUNDE, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *J. Appl. Econ.* **20**, 873–889.

HANSEN, P.R., LUNDE, A. and NASON, J.M. (2003). Choosing the best volatility models: the model confidence set approach. *Oxf. Bull. Econ. Stat.* **65**, 839–861.

HANSEN, P.R., HUANG, Z. and SHEK, H.H. (2012). Realized GARCH: a joint model for returns and realized measures of volatility. *J. Appl. Econ.* **27**, 877–906.

HARRIS, L. (2003). *Trading and exchanges: market microstructure for practitioners*. Oxford University Press, Oxford.

HARVEY, A., RUIZ, E. and SHEPHARD, N. (1994). Multivariate stochastic variance models. *Rev. Econ. Stud.* **61**, 247–264.

HARVEY, A.C. and SHEPHARD, N. (1996). Estimation of an asymmetric stochastic volatility model for asset returns. *J. Bus. Econ. Stat.* **14**, 429–434.

HASBROUCK, J. (2007). *Empirical market microstructure: the institutions, economics, and econometrics of securities trading*. Oxford University Press, Oxford.

HAUTSCH, N. (2011). Econometrics of financial high-frequency data. Springer Science & Business Media.

HAUTSCH, N., KLAUSURTAGUNG, S. and RISIKO, O. (2006). Generalized autoregressive conditional intensity models with long range dependence.

HAWKES, A.G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58**, 83–90.

HAYASHI, T. and KOIKE, Y. (2017). Multi-scale analysis of lead-lag relationships in high-frequency financial markets. arXiv:1708.03992.

HAYASHI, T., YOSHIDA, N. and ET AL (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* **11**, 359–379.

HECKMAN, J.J. and SINGER, B. (1984). Econometric duration analysis. *J. Econ.* **24**, 63–132.

HEINEN, A. (2003). Modelling time series count data: an autoregressive conditional poisson model available at SSRN 1117187.

HOFFMANN, M., ROSENBAUM, M. and YOSHIDA, N. (2013). Estimation of the lead-lag parameter from non-synchronous data. *Bernoulli* **19**, 426–461.

HOSSZEJNI, D. and KASTNER, G. (2019). Modeling univariate and multivariate stochastic volatility in R with stochvol and factorstochvol. arXiv:1906.12123.

HSIEH, M.-C., HURVICH, C. and SOULIER, P. (2019). Modeling leverage and long memory in volatility in a pure-jump process. *High Frequency* **2**, 124–141.

HUANG, D., ZHU, S., FABOZZI, F.J. and FUKUSHIMA, M. (2010). Portfolio selection under distributional uncertainty: a relative robust cvar approach. *Eur. J. Oper. Res.* **203**, 185–194.

IACUS, S.M. and YOSHIDA, N. (2018). Simulation and inference for stochastic processes with yuima. A comprehensive R framework for SDEs and other stochastic processes. Use R.

JACOD, J., LI, Y., MYKLAND, P.A., PODOLSKIJ, M. and VETTER, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stoch. Process. Appl.* **119**, 2249–2276.

JACQUIER, E., POLSON, N.G. and ROSSI, P.E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *J. Econ.* **122**, 185–212.

JASIAK, J. (1999). Persistence in intertrade durations. Available at SSRN: https://ssrn.com/abstract=162008.

JIANG, G.J. and OOMEN, R. (2005). A new test for jumps in asset prices. Preprint.

JIANG, G.J. and OOMEN, R.C. (2008). Testing for jumps when asset prices are observed with noise—a "swap variance" approach. *J. Econ.* **144**, 352–370.

KALBFLEISCH, J.D. and PRENTICE, R.L. (2011). *The statistical analysis of failure time data*. Wiley, New York.

KEIM, D.B. and MADHAVAN, A. (1996). The upstairs market for large-block transactions: analysis and measurement of price effects. *Rev. Financ. Stud.* **9**, 1–36.

KLEINBAUM, D.G. and KLEIN, M. (2010). *Survival analysis*. Springer, Berlin.

KWOK, S.S.M., LI, W.K. and YU, P.L.H. (2009). The autoregressive conditional marked duration model: statistical inference to market microstructure. *J. Data Sci.*

LANCASTER, T. (1979). Econometric methods for the duration of unemployment. *Econom. J. Econom. Soc.* **47**, 939–956.

LANE, W.R., LOONEY, S.W. and WANSLEY, J.W. (1986). An application of the cox proportional hazards model to bank failure. *J. Bank. Finance* **10**, 511–531.

LEE, S.S. and MYKLAND, P.A. (2008). Jumps in financial markets: a new nonparametric test and jump dynamics. *Rev. Financ. Stud.* **21**, 2535–2563.

LI, J., TODOROV, V., TAUCHEN, G. and LIN, H. (2019). Rank tests at jump events. *J. Bus. Econ. Stat.* **37**, 312–321.

LIBOSCHIK, T., FOKIANOS, K. and FRIED, R. (2017). tscount: an R package for analysis of count time series following generalized linear models. *J. Stat. Softw.* **82**, 1–51.

LIU, H., ZOU, J. and RAVISHANKER, N. (2018). Multiple day biclustering of high-frequency financial time series. *Stat* **7**, e176.

LIU, H., ZOU, J. and RAVISHANKER, N. (2021). Clustering high-frequency financial time series based on information theory, forthcoming. *Appl. Stoch. Models Bus. Ind.*

LIU, S. and TSE, Y.-K. (2015). Intraday value-at-Risk: an asymmetric autoregressive conditional duration approach. *J. Econ.* **189**, 437–446.

MADAN, D.B. and SENETA, E. (1990). The variance gamma (vg) model for share market returns. *J. Bus.* 511–524.

MANCINO, M.E. and SANFELICI, S. (2011). Estimating covariance via fourier method in the presence of asynchronous trading and microstructure noise. *J. Financ. Econom.* **9**, 367–408.

MANGANELLI, S. (2005). Duration, volume and volatility impact of trades. *J. Financ. Mark.* **8**, 377–399.

MARTENS, M. and VAN DIJK, D. (2007). Measuring volatility with the realized range. *J. Econ.* **138**, 181–207.

MENG, X. and TAYLOR, J.W. (2020). Estimating value-at-risk and expected shortfall using the intraday low and range data. *Eur. J. Oper. Res.* **280**, 191–202.

MIES, F., BIBINGER, M., STELAND, A. and PODOLSKIJ, M. (2020). High-frequency inference for stochastic processes with jumps of infinite activity. PhD thesis, RWTH Aachen University.

MUKHERJEE, A., PENG, W., SWANSON, N.R. and YANG, X. (2020). Financial econometrics and big data: a survey of volatility estimators and tests for the presence of jumps and co-jumps. In: Handbook of statistics, vol 42, pp 3–59. Elsevier.

NELSON, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econom. J. Econom. Soc.* **59**, 347–370.

N.Y.S.E. TRADE AND QUOTE DATABASE (2019). Retrieved from wharton research data services accessed.

O'HARA, M. (1997). *Market microstructure theory*. Wiley, New York.

PACURAR, M. (2008). Autoregressive conditional duration models in finance: a survey of the theoretical and empirical literature. *J. Econ. Surv.* **22**, 711–751.

PALMA, W. (2007). *Long-memory time series: theory and methods*. Wiley, New York.

PARKINSON, M. (1980). The extreme value method for estimating the variance of the rate of return. *J. Bus.* **53**, 61–65.

PELUSO, S., CORSI, F. and MIRA, A. (2014). A bayesian high-frequency estimator of the multivariate covariance of noisy and asynchronous returns. *J. Financ. Econom.* **13**, 665–697.

RENÒ, R. (2003). A closer look at the epps effect. *Int. J. Theor. Appl. Finance* **6**, 87–102.

ROBINSON, P.M. (2003). Time series with long memory. Advanced Texts in Econometrics.

ROCCO, M. (2014). Extreme value theory in finance: a survey. *J. Econ. Surv.* **28**, 82–108.

RUSSELL, J.R. (1999). Econometric modeling of multivariate irregularly-spaced high-frequency data. Working Paper, University of Chicago.

RYDBERG, T.H. and SHEPHARD, N. (2000). BIN models for trade-by-trade data. modelling the number of trades in a fixed interval of time. Econometric Society World Congress 2000 Contributed Papers 0740, Econometric Society. https://ideas.repec.org/p/ecm/wc2000/0740.html.

SEN, R. (2009). Jumps and microstructure noise in stock price volatility. *Volatility*, 163.

SHIROTA, S., HIZU, T. and OMORI, Y. (2014). Realized stochastic volatility with leverage and long memory. *Comput. Stat. Data Anal.* **76**, 618–641.

SO, M.K. and XU, R. (2013). Forecasting intraday volatility and value-at-risk with high-frequency data. *Asia-Pac. Finan. Markets* **20**, 83–111.

SO, M.K., CHU, A.M., LO, C.C. and IP, C.Y. (2021). Volatility and dynamic dependence modeling: review, applications, and financial risk management. *Wiley Interdiscip. Rev.: Comput. Stat.*, e1567.

SONG, X., KIM, D., YUAN, H., CUI, X., LU, Z., ZHOU, Y. and WANG, Y. (2021). Volatility analysis with realized garch-itô models. *J. Econ.* **222**, 393–410.

STROUD, J.R. and JOHANNES, M.S. (2014). Bayesian modeling and forecasting of 24-hour high-frequency volatility. *J. Am. Stat. Assoc.* **109**, 1368–1384.

SUN, W., RACHEV, S., FABOZZI, F.J. and KALEV, P.S. (2008). Fractals in trade duration: capturing long-range dependence and heavy tailedness in modeling trade duration. *Ann. Finance* **4**, 217–241.

SWISHCHUK, A. and HUFFMAN, A. (2020). General compound hawkes processes in limit order books. *Risks* **8**, 28.

TAKAHASHI, M., OMORI, Y. and WATANABE, T. (2009). Estimating stochastic volatility models using daily returns and realized volatility simultaneously. *Comput. Stat. Data Anal.* **53**, 2404–2426.

TAKAHASHI, M., WATANABE, T. and OMORI, Y. (2016). Volatility and quantile forecasts by realized stochastic volatility models with generalized hyperbolic distribution. *Int. J. Forecast.* **32**, 437–457.

TAY, A.S., TING, C., TSE, Y.K. and WARACHKA, M. (2004). Transaction-data analysis of marked durations and their implications for market microstructure.

TAY, A.S., TING, C., KUEN TSE, Y. and WARACHKA, M. (2011). The impact of transaction duration, volume and direction on price dynamics and volatility. *Quant. Finance* **11**, 447–457.

TAYLOR, S.J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961–75. *Time Ser. Anal. Theory Pract.* **1**, 203–226.

TAYLOR, S.J. (1994). Modeling stochastic volatility: a review and comparative study. *Math. Financ.* **4**, 183–204.

THAVANESWARAN, A., RAVISHANKER, N. and LIANG, Y. (2015). Generalized duration models and optimal estimation using estimating functions. *Ann. Inst. Stat. Math.* **67**, 129–156.

THERNEAU, T.M. (2021). Survival: a package for survival analysis in R. R package version 3.2-13.

TSAI, P.-C. and SHACKLETON, M.B. (2016). Detecting jumps in high-frequency prices under stochastic volatility: a review and a data-driven approach. In: Handbook of high-frequency trading and modeling in finance, pp 137–181.

TSAY, R.S. (2005). *Analysis of financial time series*. Wiley, New York.

VASILEIOS, S. (2015). acp: autoregressive conditional poisson (R package version 2.1).

WANG, Q., FIGUEROA-LÓPEZ, J.E. and KUFFNER, T.A. (2021). Bayesian inference on volatility in the presence of infinite jump activity and microstructure noise. *Electron. J. Stat.* **15**, 506–553.

WANG, Y. and ZOU, J. (2014). Volatility analysis in high-frequency financial data. *Wiley Interdiscip. Rev. Comput. Stat.* **6**, 393–404.

YAN, B. and ZIVOT, E. (2003). Analysis of high-frequency financial data with S-PLUS. Working paper, UWEC-2005-03. http://ideas.repec.org/p/udb/wpaper/uwec-2005-03.html.

YU, J. and MEYER, R. (2006). Multivariate stochastic volatility models: bayesian estimation and model comparison. *Econ. Rev.* **25**, 361–384.

ZAATOUR, R. (2014). Hawkes: Hawkes process simulation and calibration toolkit (R package version 0.0-4).

ZHANG, L. (2011). Estimating covariation: Epps effect, microstructure noise. *J. Econ.* **160**, 33–47.

ZHANG, Y., ZOU, J., RAVISHANKER, N. and THAVANESWARAN, A. (2019). Modeling financial durations using penalized estimating functions. *Comput. Stat. Data Anal.* **131**, 145–158.

ZHENG, Y., LI, Y. and LI, G. (2016). On Fréchet autoregressive conditional duration models. *J. Stat. Plan. Inference* **175**, 51–66.

ŽIKEŠ, F., BARUNÍK, J. and SHENAI, N. (2017). Modeling and forecasting persistent financial durations. *Econom. Rev.* **36**, 1081–1110.

ZIVOT, E. and WANG, J. (2007). Modeling financial time series with s-plus®, vol 191. Springer Science & Business Media.

CHIRANJIT DUTTA
NALINI RAVISHANKER
DEPARTMENT OF STATISTICS, UNIVERSITY OF CONNECTICUT, STORRS, CT, USA
E-mail: chiranjit.dutta@uconn.edu

KARA KARPMAN
DEPARTMENT OF MATHEMATICS, MIDDLEBURY COLLEGE, MIDDLEBURY, VT, USA

SUMANTA BASU
DEPARTMENT OF STATISTICS AND DATA SCIENCE, CORNELL UNIVERSITY, ITHACA, NY, USA

Paper received: 29 September 2021; accepted 22 March 2022.