EXTREME RETURNS AND INTENSITY OF TRADING [1]

WEI LIN

International School of Economics and Management
Capital University of Economics and Business, China

GLORIA GONZÁLEZ-RIVERA

Department of Economics
University of California, Riverside, USA

[1]We are grateful to the participants at the International Symposium in Forecasting, Riverside, 2015, Tsinghua International Conference on Econometrics, Beijing, 2015, and the 1st International Symposium on Interval Data Modeling, Beijing, 2015, and seminar participants at Guanghua School of Management at Peking University, School of Economics at Zhejiang University, School of Economics at Shandong University, Universidad Carlos III de Madrid (UC3M), CEMFI, Universidad Autonoma de Madrid, Universidad de Alcalá de Henares, and Universidad Complutense de Madrid for useful comments. Gloria González-Rivera wishes to thank the Department of Statistics at UC3M, where this manuscript was written, for their hospitality and for the financial support of the 2015 Chair of Excellence UC3M/Banco de Santander.

Abstract

Asymmetric information models of market microstructure claim that variables like trading intensity are proxies for latent information on the value of financial assets. We consider the interval-valued time series (ITS) of low/high returns and explore the relationship between these extreme returns and the intensity of trading. We assume that the returns (or prices) are generated by a latent process with some unknown conditional density. At each period of time, from this density, we have some random draws (trades) and the lowest and highest returns are the realized extreme observations of the latent process over the sample of draws. In this context, we propose a semiparametric model of extreme returns that exploits the results provided by extreme value theory. If properly centered and standardized extremes have well defined limiting distributions, the conditional mean of extreme returns is a highly nonlinear function of conditional moments of the latent process and of the conditional intensity of the process that governs the number of draws. We implement a two-step estimation procedure. First, we estimate parametrically the regressors that will enter into the nonlinear function, and in a second step, we estimate nonparametrically the conditional mean of extreme returns as a function of the generated regressors. Unlike current models for ITS, the proposed semiparametric model is robust to misspecification of the conditional density of the latent process. We fit several nonlinear and linear models to the 5-min and 1-min low/high returns to seven major banks and technology stocks, and find that the nonlinear specification is superior to the current linear models and that the conditional volatility of the latent process and the conditional intensity of the trading process are major drivers of the dynamics of extreme returns.

Key Words: Trading intensity, Interval-valued Time Series, Generalized Extreme Value Distribution, Nonparametric regression, Generated Regressor.

JEL Classification: C01, C14, C32, C51.

1 Introduction

Most of the financial literature has focused on the dynamics of average returns and other moments of the return distribution. We have numerous empirical studies of volatility dynamics as well as information models of market microstructure that claim that variables like trading volume (or trading intensity) are proxies for latent information on the value of financial assets (see Easley and O'Hara, 1992). Much less attention has been paid to the dynamics of extreme returns and to the links between information, proxied by trading intensity, and extreme returns.

We explore the modelling of interval-valued time series (ITS) of extreme returns, which is defined as the collection of the intervals formed by the highest and the lowest returns in a given period of time. We propose a semiparametric model that explains the generation of extreme returns as a function of volatility and trading intensity. Though a link between trading volume and volatility has already been established, the link between trading volume and extreme returns has not been analyzed in much detail. A sample of most representative results on volume and volatility follows in historical order. Lamoureux and Lastrapes (1990) find that identical latent factors drive trade volume and return volatility. Andersen (1996) proposes a model in which informational asymmetries and liquidity needs motivate trading, which in turn, drives the dynamics of a stochastic volatility model. Engle (2000) analyzes an Autoregressive Conditional Duration model and a GARCH model to conclude that the absence of trading means either bad news or no news and translates into low volatility regimes. With high frequency data (5-min intraday data), Darrat, Rahman, and Zhong (2003) find evidence of significant lead-lag relations between volume and volatility in agreement with the sequential information arrival hypothesis. Fleming and Kirby (2011) analyze the joint dynamics of trading volume and realized volatility and find that there is strong correlation between the innovations to volume and volatility. Sita and Westerholm (2011) find that trade durations (inversely related to trade volume in equity markets) have forecasting power for returns but only within the trading day. One can argue that the range of the interval of extreme returns is a very good volatility estimator (Parkinson, 1980) and in this sense, the result of the aforementioned studies may apply. However, the dynamics of the low/high interval are richer than those of the range itself because the modeling of the interval captures not only variability but also the dynamics of the bounds themselves. For instance, Ning and Wirjanto (2009) find that there is a significant and asymmetric return-volume dependence at the extremes. The largest returns tend to be associated with extremely large trading volumes but the lowest returns tend not to be related to either large or small volumes. In this context, our work offers a joint modeling of volatility, trading intensity and extreme returns with high frequency data that combines parametric and non-parametric specifications. We proceed by building up on the statistical framework that we proposed in our previous work.

The intervals formed by extreme returns have statistical properties that can be exploited. This is in contrast to the modeling of a classical time series of returns, for which it is very difficult to find any time-dependence in the average return. For instance, in González-Rivera and Lin (2013), the authors estimate a constrained bivariate linear system for the daily lowest/highest returns of the SP500 index and find that there is statistically significant dependence with adjusted R-squared (in-sample) of about 50%. Though this work generalizes specifications of previous regression models on lower/upper bounds or center/radius of intervals (see the references herein), it relies on the assumption of bivariate normality. In a subsequent analysis, unlike the regression-type models just mentioned, Lin and González-Rivera (2016) propose an alternative modeling approach by pondering how interval-valued data are generated. They consider the lower and upper bounds of the interval as the realizations of minimal and maximal order statistics coming from a sample of N_t random draws from the conditional density (at time t) of a latent random process $\{Y_t\}$. Through the statistical implementation of this approach to prices of agricultural commodities, they also find that their models provide a very good fit for extreme returns with average coverage rates (percentage overlap of the actual low/high interval with the fitted interval) of 83%. However, there are also some disadvantages of this approach. First, the joint probability density function of minimal and maximal order statistics degenerates as the number of random draws goes to infinity. Second, the normality assumption on the latent random process $\{Y_t\}$ may be too restrictive.

To overcome these drawbacks and, in particular, the restrictions imposed by the distributional assumptions, we propose a new two-step semiparametric model that exploits the extreme property of the lower and upper bounds of the interval. We maintain the general setup of Lin and González-Rivera (2016) by assuming that there is a latent process $\{Y_t\}$ with conditional density $f_{Y_t}(.)$, from which, at every moment of time, there are N_t random draws and the lower and upper bounds of the interval are the *realized extreme* observations of Y_t over the sample of draws. However, we will not assume any particular functional form of f_{Y_t}, so that the estimation procedure is robust to density misspecification of the underlying stochastic process. We will only need conditional moments of the latent process and we will rely on limiting results provided by *extreme value theory* to estimate the conditional mean of the lower and upper bounds of the interval. The proposed estimation procedure consists of two steps. First, we obtain parametric estimates of the conditional mean and conditional variance of the latent process $\{Y_t\}$ and estimates of the conditional trading intensity of the process $\{N_t\}$. Second, with the generated conditional moments of the first step as the regressors, we specify a nonparametric model for the conditional means of the lower and upper bounds. We propose a nonparametric function because, according to extreme value theory, the conditional mean of an extreme value is often a highly nonlinear function that is difficult to estimate parametrically. Thus, this semiparametric approach is a natural vehicle to analyze the role of trading intensity jointly with the statistical properties of the latent return process on the generation of extreme returns.

We fit the proposed semiparametric model to the time series of intervals of low/high returns at the 5-minute and 1-minute frequencies in the trading days of June 2017 for seven very liquid stocks: three major banks, Wells Fargo, Bank of America, and J.P. Morgan and four giant technology stocks, Amazon, Apple, Google and Intel. We find that the proposed semiparametric model is superior to the current linear specifications and there is a nonlinear relationship between extreme returns and intensity of trading with a somehow more intense response from the low returns.

The organization of this paper is the following. In section 2, we provide the basic assumptions for estimation of the model. In section 3, we present the two-step estimation procedure and establish the asymptotic properties of the second-step nonparametric regression with generated regressors. In section 4, we analyze several models to explain the relationship of extreme returns with intensity of trading and volatility, and finally, we conclude in section 5.

2 Basic Assumptions

We describe the data generating process of the interval-valued time series. We need several assumptions, which are not too restrictive, and they accommodate many of the processes frequently encountered in financial data.

Assumption 1 (Data Generating Process)**.** *Let $\{Y_t : t = 1, \cdots, T\}$ be an underlying stationary stochastic process. The continuous random variable Y_t at time t has conditional density $f(y_t|\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ is the information set available at time t. At each time t, there are N_t independent draws from $f(y_t|\mathcal{F}_{t-1})$ collected in a set $\mathcal{S}_t \equiv \{y_{it} : i = 1, \cdots, N_t\}$ with random sample size N_t, which it is assumed to follow a conditional distribution $H(n_t|\mathcal{F}_{t-1})$.*

Let y_{lt} and y_{ut} denote the smallest and largest values in the sample $\mathcal{S}_t$ at time t:

$$\begin{aligned} y_{lt} &\equiv \min_i \mathcal{S}_t = \min_{1\leq i\leq N_t}\{y_{it}\}, \\ y_{ut} &\equiv \max_i \mathcal{S}_t = \max_{1\leq i\leq N_t}\{y_{it}\}. \end{aligned}$$

Then, $\{(y_{lt}, y_{ut}) : t = 1, \cdots, T\}$ is the observed interval time series (ITS) of lower and upper bounds.

The intuition behind Assumption 1 is straightforward. For instance, suppose that we have financial data and we choose a frequency, say, every five minutes. During these five minutes, trading takes place and, for every transaction, we observe a return (price). Then, in each block of five minutes, we will observe the lowest return, the highest return, and the number of trades. Our assumption means that the conditional density of returns $f(y_t|\mathcal{F}_{t-1})$ is updated every five minutes according to some dynamic specification. The number of trades during the five-minute time interval represents the number of random draws n_t from the conditional distribution of returns. Then, the lowest and the highest returns (y_{lt} and y_{ut}) are the two extremal (minimal and maximal) observations in the sample $\mathcal{S}_t$ of size n_t.

Given this data generating mechanism, our analysis of ITS data proceeds with the analysis of extremal observations $\{(y_{lt}, y_{ut})\}$ based on the results of the extreme value theory. The asymptotic theory for maxima (and minima) is very different from the theory applied to averages. Once the average is properly centered around its mean and standardized by its standard deviation, central limit theorems provide a normal limiting distribution. In contrast, the centering and standardizing terms in the limit theorems for maxima (minima) are more difficult to derive because they depend on the tail characteristics of the assumed underlying density. The key result in extreme value theory is the Fisher-Tippett Theorem (Fisher and Tippett, 1928; Gnedenko, 1943) that provides the limiting distributions of properly centered and standardized maxima (minima) [1]. The three limiting distributions are *Fréchet*, *Weibull*, and *Gumbel*, which can be nested into a one-parameter *Generalized Extreme Value* (GEV) distribution H_ξ defined as

$$H_\xi(x) = \begin{cases} \exp\{-(1+\xi x)^{-1/\xi}\} & \text{if } \xi \neq 0, \\ \exp\{-\exp\{-x\}\} & \text{if } \xi = 0, \end{cases}$$

in which ξ is a shape parameter and $1 + \xi x > 0$. Then, (i) $\xi = \alpha^{-1} > 0$ corresponds to the Fréchet distribution, (ii) $\xi = 0$ corresponds to the Gumbel distribution, and (iii) $\xi = -\alpha^{-1} < 0$ corresponds to the Weibull distribution. For more detail, see Theorem 1.1.3 and its discussion in de Haan and Ferreira (2000).

It is said that the random variable Y_t belongs to the maximum domain of attraction (MDA) of the extreme value distribution H_ξ ($Y_t \in \text{MDA}(H_\xi)$) if the limiting distribution of standardized extremes, i.e. $c_{ut}^{-1}(Y_{ut} - d_{ut})$, is the extreme value distribution H_ξ. The standardizing and centering terms, c_{ut} and d_{ut}, depend on t through the conditional distribution of Y_t and the number of random draws N_t. Explicitly, we write $c_{ut} \equiv c_u(N_t, f(y_t|\mathcal{F}_{t-1}))$ and $d_{ut} \equiv d_u(N_t, f(y_t|\mathcal{F}_{t-1}))$. Based on Assumption 1 (DGP), the limiting distribution of maxima Y_{ut} conditioning on $\mathcal{F}_{t-1}$ is

$$c_{ut}^{-1}[Y_{ut} - d_{ut}]|\mathcal{F}_{t-1} \xrightarrow{d} H_{\xi_u}, \quad \text{for each } t = 1, \ldots, T, \tag{2.1}$$

as the number of random draws N_t goes to infinity in probability, which follows directly from the Fisher-Tippett Theorem and Lemma 2.5.6 in Embrechts, Klüppelberg, and Mikosch (1997). The same argument holds for the minima process $\{Y_{lt}\}$ so that

$$c_{lt}^{-1}[Y_{lt} - d_{lt}]|\mathcal{F}_{t-1} \xrightarrow{d} H_{\xi_l}, \quad \text{for each } t = 1, \ldots, T, \tag{2.2}$$

as the number of random draws N_t goes to infinity in probability.

A key requirement to invoke the Fisher-Tippett Theorem is that the maxima Y_{ut} and minima Y_{lt} are drawn from an i.i.d. random sample $\mathcal{S}_t$ as stated in Assumption 1. However, under certain regularity conditions, the i.i.d. assumption can be substantially relaxed to strictly stationarity, which allows the y_{it} sequence in $\mathcal{S}_t$ to be weakly dependent without essentially affecting our model

[1] We only consider continuous random variables, therefore the existence of a non-degenerate limiting distribution should always hold.

specification. We further discuss the regularity conditions in Appendix A.1.

Since we would like to build conditional mean models for the extremes, the above convergence in distribution, (2.1) and (2.2), is too weak. We need to impose restrictions on the first moments of Y_{lt} and Y_{ut} to achieve stronger convergence. For notational simplicity, let $\widetilde{Y}_{lt}(N_t) \equiv c_{lt}^{-1}(Y_{lt} - d_{lt})$ and $\widetilde{Y}_{ut}(N_t) \equiv c_{ut}^{-1}(Y_{ut} - d_{ut})$ be the appropriately standardized maxima and minima, whose dependence on the number of random draws N_t is explicitly expressed.

Assumption 2. *For all t, there exists $\delta > 0$, such that*

$$\sup_n E\left[\left|\widetilde{Y}_{lt}(n)\right|^{1+\delta}\middle|\mathcal{F}_{t-1}\right] = M_l < \infty, \quad \sup_n E\left[\left|\widetilde{Y}_{ut}(n)\right|^{1+\delta}\middle|\mathcal{F}_{t-1}\right] = M_u < \infty.$$

Given Assumption 2 and according to Theorem 4.5.2 in Chung (2001), we have that for each $t = 1, \ldots, T$,

$$\mathbb{E}\left[\widetilde{Y}_{lt}(N_t)\middle|\mathcal{F}_{t-1}\right] \xrightarrow{p} \mathbb{E}(Y_{\xi_l}), \quad \mathbb{E}\left[\widetilde{Y}_{ut}(N_t)\middle|\mathcal{F}_{t-1}\right] \xrightarrow{p} \mathbb{E}(Y_{\xi_u}),$$

as N_t goes to infinity in probability. Since the conditional expectation of the GEV random variable Y_{ξ_u} is $E(Y_{\xi_u}) = [\Gamma(1-\xi_u) - 1]/\xi_u$, where $\Gamma(\cdot)$ is the Gamma function, the conditional expectations of the extrema are

$$E(Y_{ut}|N_t;\mathcal{F}_{t-1}) = d_u(N_t, f(y_t|\mathcal{F}_{t-1})) + c_u(N_t, f(y_t|\mathcal{F}_{t-1}))\frac{\Gamma(1-\xi_u)-1}{\xi_u} + o(c_u(N_t,\theta_t)),$$
$$E(Y_{lt}|N_t;\mathcal{F}_{t-1}) = d_l(N_t, f(y_t|\mathcal{F}_{t-1})) + c_l(N_t, f(y_t|\mathcal{F}_{t-1}))\frac{\Gamma(1-\xi_l)-1}{\xi_l} + o(c_l(N_t,\theta_t)).$$

The conditional mean functions of the upper and lower bounds depend on the centering and standardizing terms associated with the assumed conditional density $f(y_t|\mathcal{F}_{t-1})$. Even for some common densities like normal or Student's t, these terms are highly nonlinear on the moments of interest.[2] Therefore, we propose to estimate the conditional mean functions nonparametrically so that they are robust to density misspecification of the underlying stochastic processes. In doing

[2] If y_t is normally distributed $N(\mu_t, \sigma_t^2)$, we have

$$c_u(n_t, \mu_t, \sigma_t) = \frac{1}{\sigma_t\sqrt{2\ln n_t}}; \; d_u(n_t, \mu_t, \sigma_t) = \mu_t + \sigma_t\sqrt{2\ln n_t} - \sigma_t\frac{\ln(4\pi) + \ln\ln n_t}{2(2\ln n_t)^{1/2}}.$$

If y_t has t-distribution with mean μ_t and degrees of freedom ν_t, we have $d_u(n_t, \mu_t, \sigma_t) = 0$ and $c_u(n_t, \mu_t, \nu_t)$ is the solution to the following reduced form model,

$$\frac{1}{n_t} = \frac{1}{2} - (c-\mu_t)\Gamma\left(\frac{\nu_t+1}{2}\right)\cdot\frac{{}_2F_1\left(\frac{1}{2}; \frac{\nu_t+1}{2}; \frac{3}{2}; -\frac{(c-\mu_t)^2}{\nu_t}\right)}{\sqrt{\pi\nu_t}\Gamma\left(\frac{\nu_t}{2}\right)}$$

where ${}_2F_1$ is the hypergeometric function.

so, we also avoid the difficult task of calculating the associated standardizing and centering terms. Then, we write

$$E(Y_{ut}|N_t, \mathcal{F}_{t-1}) = m_u(N_t, f(y_t|\mathcal{F}_{t-1}), \xi_u), \tag{2.3}$$

$$E(Y_{lt}|N_t, \mathcal{F}_{t-1}) = m_l(N_t, f(y_t|\mathcal{F}_{t-1}), \xi_l), \tag{2.4}$$

where $m_l(\cdot)$ and $m_u(\cdot)$ are the conditional mean functions depending on the conditional density of the underlying process $f(y_t|\mathcal{F}_{t-1})$, the number of random draws N_t, and the shape parameters of the limiting GEV distribution ξ_l and ξ_u. Note that ξ_l and ξ_u are constant, and therefore can be innocuously excluded from the functions.

We also assume that the conditional density $f(y_t|\mathcal{F}_{t-1})$ is indexed by a finite-dimensional parameter. We will include the first two moments of the underlying random process Y_t in a parameter vector θ_t, i.e., $\theta_t = (\mu_t, \sigma_t)$, to capture the location and the scale of the conditional distribution of Y_t at time t. Similarly, for the number of random draws N_t, we assume that the conditional distribution $H(n_t|\mathcal{F}_{t-1})$ is indexed by the first moment of N_t, denoted by λ_t. Formally,

Assumption 3. *(i) For any time period t, the conditional density $f(y_t|\mathcal{F}_{t-1})$ is indexed by the first and second order conditional moments $\theta_t \equiv \theta(\mathcal{F}_{t-1}) \in \Theta \subset \mathbb{R}^2$, where Θ is a compact subset of the Euclidean space, i.e., $f(y_t|\mathcal{F}_{t-1}) = f(y_t; \theta_t)$ for all t.*

(ii) For any time period t, the conditional distribution $H(n_t|\mathcal{F}_{t-1})$ is indexed by the first order conditional moment $\lambda_t \equiv \lambda(\mathcal{F}_{t-1}) \in \Theta \subset \mathbb{R}$, where Θ is a compact subset of the Euclidean space, i.e., $H(n_t|\mathcal{F}_{t-1}) = H(n_t; \lambda_t)$ for all t.

(iii) Let Ψ_1 and Ψ_2 be compact subsets on some finite k-dimensional Euclidean space $\mathbb{R}^k$. The expectational models $\mathcal{M}_1(\Psi_1)$ and $\mathcal{M}_2(\Psi_2)$ are correctly specified for $\theta_t \equiv (\mu_t, \sigma_t^2)$ and λ_t, respectively, i.e.,

$$\begin{aligned}
\mu_t &\equiv E(Y_t|\mathcal{F}_{t-1}) = \mu(\mathcal{F}_{t-1}, \psi_1^o) \\
\sigma_t^2 &\equiv E[(Y_t - \mu_t)^2|\mathcal{F}_{t-1}] = \sigma^2(\mathcal{F}_{t-1}, \psi_1^o) \\
\lambda_t &\equiv E(n_t|\mathcal{F}_{t-1}) = \lambda(\mathcal{F}_{t-1}, \psi_2^o)
\end{aligned}$$

almost surely for each time t with some $\psi_1^o \in \Psi_1$ and $\psi_2^o \in \Psi_2$. In addition, the point-valued *time series $\{y_t\}_{t=1}^T$ (e.g. returns based on closing prices), and $\{n_t\}_{t=1}^T$, used to estimate the parameters in the specifications $\mathcal{M}_1$ and $\mathcal{M}_2$, satisfy regularity conditions such that the estimates $\widehat{\psi}_1$ and $\widehat{\psi}_2$ are $\sqrt{T}$-consistent.*

Given assumption 3(i), the conditional expectations of maxima and minima in (2.3) and (2.4) can be further simplified to

$$E(Y_{lt}|N_t, \mathcal{F}_{t-1}) = m_l(N_t, \theta_t), \quad E(Y_{ut}|N_t, \mathcal{F}_{t-1}) = m_u(N_t, \theta_t). \tag{2.5}$$

For these conditional expectations, observe that the conditioning information set includes not only past information $\mathcal{F}_{t-1}$ but also the number of random draws N_t in the current period. Since N_t is not observable until time period t ends, econometric models directly built upon (2.5) cannot be used to forecast future extreme returns in practice. To overcome this drawback, we integrate out the random variable N_t in (2.5) so that the conditioning information set only contains $\mathcal{F}_{t-1}$, which is available at the beginning of time t. With assumption 3(ii), we can calculate the marginal expectations of the extremes as

$$E(Y_{lt}|\mathcal{F}_{t-1}) = M_l(\theta_t, \lambda_t) \equiv \int m_l(s, \theta_t) dH(s; \lambda_t), \tag{2.6}$$

$$E(Y_{ut}|\mathcal{F}_{t-1}) = M_u(\theta_t, \lambda_t) \equiv \int m_u(s, \theta_t) dH(s; \lambda_t). \tag{2.7}$$

Assumption 3(iii) is a high level assumption. In the framework of QMLE, it requires that the quasi log-likelihood function obeys the strong uniform law of large numbers (SULLN). Primitive conditions are available in the literature, see Domowitz and White (1982), among others.

3 Estimation

We propose to estimate (2.6) and (2.7) in two steps. First, we will generate the regressors θ_t and λ_t, and secondly we will estimate non-parametrically the conditional mean functions.

If the parameter λ_t and θ_t were known, we could directly use nonparametric methods to estimate the following two conditional mean models:

$$Y_{lt} = M_l(\theta_t, \lambda_t) + \varepsilon_{lt}, \tag{3.1}$$

$$Y_{ut} = M_u(\theta_t, \lambda_t) + \varepsilon_{ut}. \tag{3.2}$$

However, in most situations the regressors λ_t and θ_t are unknown. We will estimate them by proposing some parametric models that, according to assumption 3(iii), must be well specified. Consequently, our objective is the estimation of nonparametric conditional mean functions of generated regressors:

$$Y_{lt} = M_l(\widehat{\theta}_t, \widehat{\lambda}_t) + v_{lt}, \tag{3.3}$$

$$Y_{ut} = M_u(\widehat{\theta}_t, \widehat{\lambda}_t) + v_{ut}. \tag{3.4}$$

To estimate $\theta_t \equiv (\mu_t, \sigma_t^2)$, we work with a point-valued time series. If we are modelling returns, we can follow the standard practice of choosing the series of returns calculated at the end of each time period. Alternatively, we could also choose the series of the centers of the intervals as a realized sample path of the underlying process $\{Y_t\}$ and specify the dynamics of the conditional mean of the centers. The specification of the dynamics of the variance could be based on the time series of ranges of the intervals. Similarly, we work with the realized sample path $\{n_t\}$, specify and estimate the dynamics of the conditional intensity $\lambda_t = E(N_t|\mathcal{F}_{t-1})$.

It is possible to avoid these generated regressors by directly inserting into the nonparametric functions those observed regressors in the information set $\mathcal{F}_{t-1}$ that drive the conditional moments μ_t, σ_t^2 and λ_t. The drawback of this approach is that the number of regressors could be very large so we face the curse of dimensionality of nonparametric models. The generated regressor approach offers a more parsimonious model, though we need to take into account the extra uncertainty generated by the estimation of the regressors.

There are two important differences with the approach in Lin and González-Rivera (2016). There, the estimation methodology is maximum likelihood and the log-likelihood function is based on the joint density of the lowest and the highest rank order statistics of the random sample $\mathcal{S}_t \equiv \{y_{it} : i = 1, \cdots, N_t\}$. Though we assume conditional normality for the underlying process $\{Y_t\}$, a QML estimator may not exist because, as we discussed there, the joint density of the ordinal statistics does not belong to the quadratic exponential family and the consistency of the QML estimator cannot be guaranteed. The approach that we propose here is robust to distributional assumptions:

- With the realized sample paths of point-valued time series, i.e., $\{y_t\}$ and $\{n_t\}$, associated with the underlying stochastic processes $\{Y_t\}$ and $\{N_t\}$, we estimate consistently the conditional moments (θ_t, λ_t).
- With the maxima and minima of the interval-valued time series $(\{y_{lt}\}, \{y_{ut}\})$ and the parametrically generated covariates $(\widehat{\theta}_t, \widehat{\lambda}_t)$, we estimate nonparametrically the two conditional mean functions (3.3) and (3.4).

The second difference is related to the feasibility of the order statistics approach when the number of draws N_t is very large. A quick look at the log-likelihood function based on the joint density of the ordinal statistics reveals that, for large number of trades, the function will explode, and the optimization exercise will not have a solution. Hence, the extreme value approach proposed here is general enough to provide both, time feasibility and robustness.

Under Assumption 3(iii) and within a QMLE framework, the first-step estimators $(\widehat{\theta}_t, \widehat{\lambda}_t)$ enjoy standard asymptotic properties. Now, we focus on the asymptotics of the second step nonparametric estimator. In this respect, we follow Conrad and Mammen (2008) who estimate a semiparametric GARCH-in-Mean model in which the dynamics of the conditional variance are parametrically specified, and the dependence of the return on its conditional variance is estimated by nonparametric kernel smoothing methods. Our two-step estimator is similar to their iterated estimator but much simpler. First, the latent regressors in our model are generated parametrically, while their first step estimators have both parametric and nonparametric components. Second, because the conditional variance enters the nonparametric conditional mean function and depends on past error terms, the estimator in Conrad and Mammen (2008) need an iterative estimation procedure. In contrast, our first-step estimates are obtained from parametric models based on realized sample paths of the underlying process. Hence, an iterative estimation procedure is not needed for our two-step estimator. The only difficulty is that our nonparametric conditional mean functions involve multiple generated covariates as opposed to a single covariate in Conrad and Mammen (2008). Therefore, we only need a mild adaptation of their theorems to show that the oracle property of a kernel-based nonparametric estimator also applies to our two-step estimator.

Before stating Theorem 1, we first introduce some terms to simplify notation. Let $h_t \equiv h_t(\psi_0) = (\theta_t(\psi_0), \lambda_t(\psi_0))$ be the finite q-dimensional random process of true moments and $\widehat{h}_t \equiv h_t(\widehat{\psi})$ be their estimates obtained in the first step estimation. Let $\sigma_j^2(x) = \mathbb{E}(\varepsilon_{jt}^2 | h_t = x)$ be the conditional variance of the error terms ε_{jt} ($j = l$ and u) in (3.1) and (3.2). Let $f_h(x)$ be the q-dimensional unconditional PDF of h_t.

Theorem 1 (Asymptotic properties of the two-step local linear estimator)**.**
For $j = l$ and u, assume that $M_j(x)$, $f_h(x)$, and $\sigma_j^2(x)$ are twice differentiable. $K(v) = \prod_{\ell=1}^q k(v_\ell)$ is a bounded second order kernel, $\kappa_{2,0} = \int K(v)^2 dv$, $\kappa_{1,2} = \int K(v) v^2 dv$, and $b = (b_1, \cdots, b_q)$ are the bandwidths for the q variables in h_t with $Tb_1 \cdots b_q \to \infty$. Given Assumptions 1 – 3 and the regularity conditions stated in Appendix S.1 of the supplementary file, we have

(i) (Asymptotic Equivalence) For $j = l$ and u, the two-step local linear estimator $\widehat{M}_j^{LL}(x)$ with generated covariates $\widehat{h}_t$ is asymptotic equivalent to the infeasible estimator $\widetilde{M}_j^{LL}(x)$ in the sense that

$$\sup_{x \in I} |\widehat{M}_j^{LL}(x) - \widetilde{M}_j^{LL}(x)| = o_p(T^{(\eta_+ - 1)/2} + T^{-2\eta_+}),$$

where $\eta_+ = \sum_{\ell=1}^q \eta_\ell$, and the two terms $T^{(\eta_+ - 1)/2}$ and $T^{-2\eta_+}$ are the orders of the leading variance and bias terms for the infeasible estimator $\widetilde{M}_j^{LL}(x)$ respectively;

(ii) (Asymptotic Normality) For $j = l$ and u, the limiting distribution of the feasible two-step local linear estimator is the same as that of the infeasible estimator, i.e.,

$$\sqrt{Tb_1 \cdots b_q} \left(\widehat{M}_j^{LL}(x) - M_j(x) - \frac{\kappa_{1,2}}{2} \sum_{\ell=1}^q b_\ell^2 M_j^{(2)}(x) \right) \xrightarrow{d} N\left(0, \frac{\kappa_{2,0}^q \sigma_j^2(x)}{f_h(x)}\right).$$

We provide a sketched proof of Theorem 1 in Appendix S.2 of the supplementary file. Essentially, the regularity conditions require that (a) the process from which the true moments h_t will be estimated be stationary and β-mixing; (b) the estimates $\hat{h}_t$ converge to their true values at the $\sqrt{T}$-rate, which is fast enough; (c) the dynamic specifications in Assumption 3(iii) can be well approximated by linear functions of parameters ψ in a neighborhood of their true values ψ^0; (d) the error terms ε_{lt} and ε_{ut} have conditional subexponential tails; and (e) some other regularity conditions on the kernel functions controlling the bias terms of the local linear smoothing.

4 Extreme Returns and Intensity of Trading

We model the interval-valued time series of the low/high returns to seven stocks in the U.S. financial and technology sectors. The financial stocks correspond to three banks: Wells Fargo Corporation (WFC), Bank of America (BAC), and JP Morgan Chase (JPM), which are traded in the New York Stock Exchange. The technology stocks correspond to the four tech giants: Apple (AAPL), Amazon (AMZN), Google (GOOG), and Intel (INTC) , which are traded in the Nasdaq Stock Market. All these stocks are very liquid as their trading volumes are very high. We analyze the time series at the 5-minute and 1-minute frequencies from June 1st to June 30th, 2017 for a total of 22 trading days. The data are retrieved from the NYSE Trade and Quote (TAQ) database that provides all trades and quotes occurred on the trading days of June 2017 for all those stocks. We record the stock price of every trade and split the trading day in 5-minute and 1-minute periods so that for each stock we have two time series to model. We compute the log-returns with respect to the last price of the previous period and report the returns in basis points (one basis point is defined as one per ten-thousand, i.e. 1 basis point = 1‱):

$$\begin{aligned} y_{ut} &\equiv \log(P_{high,t}/P_{close,t-1}) \times 10,000‱, \text{(the highest return)} \\ y_{lt} &\equiv \log(P_{low,t}/P_{close,t-1}) \times 10,000‱, \text{(the lowest return)} \\ y_{ct} &\equiv \log(P_{close,t}/P_{close,t-1}) \times 10,000‱, \text{(close-to-close return)} \end{aligned}$$

where $P_{high,t}$ and $P_{low,t}$ are the highest and lowest price in the period t, and $P_{close,t-1}$ is the last price in the previous period $(t-1)$. Given the interval-valued return $[y_{lt}, y_{ut}]$, the center c_t and range r_t are defined as

$$\begin{aligned} c_t &\equiv \frac{y_{lt} + y_{lt}}{2} = \log(\sqrt{P_{high,t}P_{low,t}}/P_{close,t-1}) \times 10,000‱, \\ r_t &\equiv y_{ut} - y_{lt} = \log(P_{high,t}/P_{low,t}) \times 10,000‱. \end{aligned}$$

The total number of observations is 1716 (22 days × 78 observations per day) at the 5-minute frequency; and 8580 (22 days × 390 observations per day) at the 1-minute frequency.

The complexity of the estimation generates a sheer number of models and, due to space constraints, we will only report a subset of results [3]. We showcase our modelling strategy for one bank BAC and for one technology stock AMZN and provide an additional file with supplementary material on the description and estimation results of the five remaining stocks.

Prior to the analysis, we analyze the outliers in the samples. We implement a modified version of Tukey's fences (Tukey, 1977) for identification and removal of outliers that is more conservative than that provided by Brownless and Gallo (2006) and Barndorff-Nielsen et al. (2009). The detailed procedure is explained in the Appendix A.2.

In Table 1, we start by reporting descriptive statistics for BAC and AMZN lowest and highest returns as well as the close-to-close returns and the time series of the number of trades at both 5-minute and 1-minute frequencies. We also report the characteristics of the time series of the center and range of the low/high interval of returns. These series are also plotted in Figure 1.

[Table 1] [Figure 1]

As expected, the highest returns are positively skewed and the lowest returns negatively skewed, both with large kurtosis. There is a mild positive correlation or no correlation between the highest and the lowest returns, and a mild negative correlation between the center and range series. The number of trades series exhibit overdispersion with a variance much larger than the mean favoring a negative binomial distribution. The 1-minute time series are much less volatile than the 5-minute series, which is expected. On average, the number of trades at the 5-minute frequency is about five times the number of trades at the 1-minute frequency. For BAC, the close-to-close returns and the center series are very similar, and they seem to be symmetric around a median (or mean) of zero with no much skewness and mild kurtosis. The range series is positively skewed with large kurtosis. These features hold at both frequencies and are also present in the JPM and WFC series (see Tables S3 and S4 in the supplementary file). For AMZN, we observe similar features to those in BAC but more pronounced; there is more volatility, more skewness and more kurtosis in the AMZN and other tech return series than those in the banking time series. A distinctive feature for AMZN and the

[3] The first step estimation involves searching for the best models for the conditional mean and the conditional variance of the latent process, and for the conditional intensity of the number of trades. With seven stocks and two frequencies, we end up with 42 final models ($7 \times 2 \times 3$). In the second step estimation, we search for the best nonparametric specification for low and high returns. We entertain six nonparametric models and two linear models. With seven stocks and two frequencies, we end up with 112 final models ($7 \times 2 \times 8$).

rest of the technology stock returns is that the lowest returns tend to be more volatile and more skewed than the highest returns. Even after removing outliers, the lowest returns tend to be much larger in magnitude than the highest returns (see Tables S5, S6 and S7 in the supplementary file).

4.1 Conditional moments of the latent process

We proceed to model the conditional moments of the underlying latent process Y_t. We implement two different approaches to model the dynamics of Y_t. First, we consider the center and range of the low/high intervals as good proxies for the location and scale of the conditional probability density function of Y_t. We will call this approach the "interval value approach for Y_t". Second, we simply use the close-to-close returns series $\{y_{ct}\}$ as a realized sample path of the underlying latent processes Y_t and model its conditional moments. We will call this approach the "point value approach for Y_t". Due to the space constraints, we report the estimation results for the interval value approach here and those for the point value approach in the supplementary file (see Section S.3 in the supplementary file).

We model the dynamics of the center and range series separately. In all our specification searches, we select the best model, i.e., optimal number of lags, by minimizing AIC. This criterium is rather conservative as it tends to choose models with a large number of lags. In our modelling strategy, we prefer to be conservative in the first step estimation to guard against potential misspecification. As we have described in the previous sections, the estimated conditional moments of the latent process are inputs into the final nonparametric models of the low/high returns and they need to be correctly specified for the results of Theorem 1 to go through. Thus, we prefer less parsimonious models, even at the cost of carrying some noise into the second step estimation. In addition, each specification needs to pass standard diagnosis tests, i.e. the residuals must be white noise and the pseudo Pearson residuals must have zero mean and unit standard deviation.

For the center series $\{c_t\}$, we fit simple ARMA models. The preferred specifications for the center series of BAC are ARMA(3,1) (5-min frequency) and MA(2) (1-min frequency). For AMZN, AR(1) (5-min frequency) and ARMA(1,2) (1-min frequency). In the top panel of Table 2, we report the estimations results of the ARMA models for the center series of BAC and AMZN at 5-minute frequency (see Table S9 in the supplementary file for results at the 1-minute frequency). In Figure 2, we plot the fitted time series versus the actual series.

[Table 2] [Figure 2]

As expected, the time dependence in the conditional mean is very weak in both stocks and it only reflects some microstructure noise. In the case of BAC, the mean is zero and in the case of AMZN the mean is negative. In comparison with the actual values of the center time series, the conditional means are practically zero. This is evident in the time series plots of the center series in Figure 2. We will call $\widehat{c}_t$ the fitted value for the center that proxies the estimated conditional mean of the latent process.

For the range series $\{r_t\}$, we specify a Conditional AutoRegressive Range model with Burr distribution (CARR-Burr). Since the original range series $\{r_t\}$ exhibit strong diurnal patterns, we first remove the intraday seasonality for each weekday separately by cubic B-spline smoothing, that is,

$$r^*_{t(d)} = \frac{r_{t(d)}}{f_d(i_{t(d)})}, \quad f_d(i_{t(d)}) = \exp\left(b_0 + \sum_{j_d=1}^{L_d} b_{j_d} B_{j_d}(i_{t(d)})\right), \tag{4.1}$$

where $\{B_{j_d}(\cdot) : j_d = 1, \ldots, L_d\}$ are B-spline basis functions, $t(d)$ selects those observations on weekday $d \in \{\text{Monday, Tuesday, Wednesday, Thursday, Friday}\}$, and $i_{t(d)}$ is the fraction of time in the trading day for the t-th observation such that

$$i_{t(d)} = \begin{cases} 1, & \text{if } t(d) \bmod D = 0 \\ t(d)/D - \lfloor t(d)/D \rfloor, & \text{otherwise,} \end{cases}$$

where D is the total number of observations in each day. We have $D = 78$ for the series in 5-minute frequency and $D = 390$ for the series in 1-minute frequency. After taking natural logarithm on both sides of (4.1), we obtain the coefficient estimates $\widehat{b}_j$ and $\widehat{f}_d(i_{t(d)})$ by ordinary least squares and the number of splines L_d is selected by generalized cross-validation (reported in Table S8 of the supplementary file); and the estimated intraday seasonality for each time period t is $\widehat{f}(t) = \widehat{f}_d(i_{t(d)})$ if the time period t is in weekday d.

Then, we specify the conditional autoregressive range model with Burr distribution, CARR(p, s)-

Burr, for the adjusted range r_t^*,

$$\begin{aligned} r_t^* &= \psi_t \varepsilon_t, \\ \psi_t &= \omega + \sum_{i=1}^{p} \alpha_i r_{t-i}^* + \sum_{j=1}^{s} \beta_j \psi_{t-j}, \\ \varepsilon_t | \mathcal{F}_{t-1} &\sim g_{\text{Burr}}(\cdot; \theta), \end{aligned}$$

where ψ_t is the conditional mean of the adjusted range based on the information set available at time t. The normalized adjusted range $\varepsilon_t = r_t^*/\psi_t$ is assumed to be standardized Burr distributed with density function $g_{\text{Burr}}(\cdot; \theta)$ with unit mean and shape parameters $\theta \equiv (\kappa, \sigma^2)$. We impose the restriction $\sum_{i=1}^{r} \alpha_i + \sum_{j=1}^{s} \beta_j < 1$ to ensure that the series r_t^* is stationary. As with the center series, the optimal lags p and s in CARR(p, s)-Burr are selected by AIC and the adequacy of the models is assessed with standard diagnostics. The selected model for the range series of BAC is CARR(3,4) (5-min frequency) and CARR(5,6) (1-min frequency). For AMZN, CARR(6,5) (5-min frequency) and CARR(9,10) (1-min frequency).

We denote $\widehat{r}_t^*$ as the fitted range with diurnal pattern adjustment and $\widehat{r}_t = \widehat{r}_t^* \widehat{f}(i_t)$ the fitted range for the original range. In the bottom panel of Table 2, we report the estimation results of the CARR-Burr models for the range series of BAC and AMZN at 5-minute frequency (see Table S8 for results at the 1-minute frequency). Both series have large persistence, for BAC the persistence is about 0.82 (obtained by $\sum \alpha_i + \sum \beta_j$), while that of AMZN is 0.95. The shape parameters κ and σ^2 are significantly different from 1 and 0 respectively for both stocks so that the empirical conditional standardized probability density is far from exponential. Standard diagnostic tests indicate that the fitting is adequate. The standardized residuals have zero mean and unit standard deviation. The Q-statistics of the standardized residuals show no residual dependence left in the data. We plot the fitted range series in Figure 2. For the remaining five stocks, the preferred specifications for the center and range series are in Table S11 of the supplementary file.

4.2 Conditional trading intensity

We specify autoregressive dynamics in the conditional trading intensity to account for the temporal dependence in the number of trades series $\{n_t\}$. As in the range series $\{r_t\}$, the number of trades series exhibits a clear diurnal pattern. We remove the intraday seasonality for each weekday separately by spline smoothing on the original series, that is,

$$n_t^* = n_t / f_d(i_{t(d)}), \tag{4.2}$$

where the intraday seasonality $f_d(i_{t(d)})$ is defined and obtained as in the case of the range series explained in Section 4.1. Then, for the adjusted numbers of trades n_t^*, we specify the model

$$\psi_t = \omega + \sum_{i=1}^{p} \alpha_i n_{t-i}^* + \sum_{j=1}^{s} \beta_j \psi_{t-j}. \tag{4.3}$$

Combining (4.2) and (4.3), we propose the Autoregressive Conditional Intensity (ACI) model,

$$\lambda_t = \widehat{f}(i_t)\psi_t, \quad \text{and} \quad n_t \sim g_{\text{NB}}(\cdot; \lambda_t, d), \tag{4.4}$$

where λ_t is the conditional trading intensity based on the information set available at time t. The number of trades N_t is assumed to be negative binomial distributed with density function $g_{\text{NB}}(\cdot; \lambda_t, d)$ with mean λ_t and dispersion parameter d. We restrict $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{s} \beta_j < 1$ to ensure that the series $\{n_t\}$ is stationary. We select the optimal number of lags p and s by AIC and test the specification of the model with standard diagnostic statistics. The estimated conditional trading intensity is denoted as $\widehat{\lambda}_t$.

For the BAC series, the preferred models are ACI(12,13) (5-min frequency) and ACI(12,12) (1-min frequency). For the AMZN series, they are ACI(10,10) (5-min and 1-min frequencies). We report the estimation results of the ACI models for the number of trades series of BAC and AMZN at 5-minute frequency in Table 3 (see Table S10 for the results at the 1-minute frequency) and we plot the estimated conditional intensity against the actual number of trades in Figure 2. Both series have a strong persistence, the persistence is 0.75 for BAC and 0.81 for AMZN. The pseudo-Pearson residuals have mean zero and variance one and their Q-statistics do not show any dependence, indicating that these specifications are adequate. For the remaining five stocks, the preferred specifications for the number of trades series are displayed in Table S11 of the supplementary file.

[Table 3] [Figure 2]

4.3 Nonparametric conditional mean for lower and upper bounds

From the modeling of the latent process Y_t, we gather the estimated regressors, i.e., conditional mean, range, and intensity, that will be fed into a nonparametric regression to obtain the conditional means of the lower and upper bounds, i.e., y_{lt} and y_{ut}, of the return interval. We propose and evaluate the following nonparametric regressions.

The first model has as regressors the estimated conditional centers $\widehat{c}_t$ (ARMA models), conditional mean ranges $\widehat{r}_t$ (CARR-Burr models), and conditional intensity $\widehat{\lambda}_t$ (ACI-NB models). It is the most general and parsimonious model.

- **Model 1**: $y_{jt} = M_j(\widehat{c}_t, \widehat{r}_t, \widehat{\lambda}_t) + v_{jt}, \quad \text{for } j = l, u.$

If the models for centers, ranges, and intensity have short dynamics, we could avoid the first estimation step and directly include original regressors such as c_{t-1}, r_{t-1}, n_{t-1}, etc. into the nonparametric regressions. The drawback of this approach is that, if the number of lags is very large, we run into the curse of dimensionality. With the current data, we experiment with the following model,

- **Model 2**: $y_{jt} = M_j(c_{t-1}, r_{t-1}, n_{t-1}) + v_{jt}, \quad \text{for } j = l, u.$

The next model considers regressors when the modelling of Y_t follows the "point value approach". We estimate the conditional standard deviation $\widehat{\sigma}_t$ (GARCH-GED models) (see Tables S1 and S2 in the supplementary file) in addition to the conditional intensity $\widehat{\lambda}_t$ (ACI models), $\widehat{\lambda}_t$, i.e.,

- **Model 3**: $y_{jt} = M_j(\widehat{\sigma}_t, \widehat{\lambda}_t) + v_{jt}, \quad \text{for } j = l, u.$

We propose the next two nonparametric models to assess the importance of the intensity of trading in the modelling of the conditional mean of the upper and lower bounds,

- **Model 4**: $y_{jt} = M_j(\widehat{c}_t, \widehat{r}_t) + v_{jt}, \quad \text{for } j = l, u.$

- **Model 5**: $y_{jt} = M_j(c_{t-1}, r_{t-1}) + v_{jt}, \quad \text{for } j = l, u.$

Finally, we also propose Model 6 to assess whether the conditional trading intensity $\widehat{\lambda}_t$ could be sufficient to predict the expected lower and upper bounds without information about the latent process Y_t,

- **Model 6**: $y_{jt} = M_j(\widehat{\lambda}_t) + v_{jt}, \quad \text{for } j = l, u.$

4.4 In-sample model evaluation

We evaluate the performance of the proposed models by comparing several measures of fit for interval-valued data. In addition to the six specifications of the previous section, we include two additional models proposed in González-Rivera and Lin (2013), which are constrained VAR-type specifications satisfying the inequality $y_{lt} \leq y_{ut}$ for all t. These are Interval Autoregressive-Two Step (IAR-TS) and Interval Autoregressive-Modified Two Step (IAR-MTS). These models have been proven to be superior to the existing interval-valued regression approaches (see González-Rivera and Lin, 2013; Lin and González-Rivera, 2016).

For a sample of size T, let $[\widehat{y}_{lt}, \widehat{y}_{ut}]$ be the fitted values of the corresponding interval $\mathbf{y}_t = [y_{lt}, y_{ut}]$ provided by each model. We consider the following criteria:

(i) Mean Squared Error (MSE) for upper and lower bounds separately.

$$\text{MSE}_{\text{lower}} = \sum_{t=1}^{T} (\widehat{y}_{lt} - y_{lt})^2 / T \qquad \text{MSE}_{\text{upper}} = \sum_{t=1}^{T} (\widehat{y}_{ut} - y_{ut})^2 / T;$$

(ii) Multivariate Loss Functions (MLF) for the vector of lower and upper bounds (Komunjer and Owyang, 2012): $L_p(\boldsymbol{\tau}, \mathbf{e}) \equiv (\|\mathbf{e}\|_p + \boldsymbol{\tau}'\mathbf{e}) \, \|\mathbf{e}\|_p^{p-1}$ where $\| \cdot \|_p$ is the l_p-norm, $\boldsymbol{\tau}$ is a two-dimensional parameter vector that determines the asymmetry of the loss function (if $\boldsymbol{\tau} = \mathbf{0}$, the bivariate loss is symmetric), and $\mathbf{e} = (e_l, e_u)$ is the bivariate residual interval $(\widehat{y}_{lt} - y_{lt}, \widehat{y}_{ut} - y_{ut})$. We consider two norms, $p = 1$ and $p = 2$ and their corresponding $\boldsymbol{\tau}$ parameter vectors within the unit balls $\mathcal{B}_\infty \equiv \{(\tau_1, \tau_2) \in \mathbb{R}^2 : |\tau_1| \leq 1 \text{ and } |\tau_2| \leq 1\}$ and $\mathcal{B}_2 \equiv \{(\tau_1, \tau_2) \in \mathbb{R}^2 : \tau_1^2 + \tau_2^2 \leq 1\}$, respectively.

Then, the Multivariate Loss Functions (MLF) are defined by their sample averages:

$$MLF_1 = \sum_{t=1}^{T} L_1(\boldsymbol{\tau}_t^*, \mathbf{e}_t)/T, \qquad MLF_2 = \sum_{t=1}^{T} L_2(\boldsymbol{\tau}_t^*, \mathbf{e}_t)/T,$$

where $\boldsymbol{\tau}_t^*$ is the optimal vector that defines the asymmetry of the loss.

(iii) Mean Distance Error (MDE) between the fitted and actual intervals (Arroyo, González-Rivera, and Mate, 2011).

Let $D_q(\widehat{y}_t, y_t)$ be a distance measure of order q between the fitted and the actual intervals, the mean distance error is defined as $MDE_q(\{\widehat{y}_t\}, \{y_t\}) = \sum_{t=1}^{T} D_q^q(\widehat{y}_t, y_t)/T$. We consider $q = 1$ and $q = 2$, with a distance measure such as,

$$
\begin{aligned}
D_1(\widehat{y}_t, y_t) &= \frac{1}{2}(|\widehat{y}_{lt} - y_{lt}| + |\widehat{y}_{ut} - y_{ut}|), \\
D_2(\widehat{y}_t, y_t) &= \frac{1}{\sqrt{2}}[(\widehat{y}_{lt} - y_{lt})^2 + (\widehat{y}_{ut} - y_{ut})^2]^{1/2}.
\end{aligned}
$$

Note that MDE_1 and MDE_2 are equal to a half of MLF_1 and MLF_2 respectively if $\boldsymbol{\tau} = \mathbf{0}$.

In Table 4, we report the in-sample evaluation of the linear and nonparametric models for BAC and AMZN stock returns at both 5-minute and 1-minute frequencies. The supplementary file contains similar tables (S12 – S16) with the evaluation results for the remaining five stocks. The first finding is that the nonparametric regressions are superior to the IAR-TS and IAR-MTS specifications as they deliver, in most cases, the smallest losses across loss functions and for both BAC and AMZN stocks at both 5-minute and 1-minute frequencies. Within the six nonparametric models, the preferred specification is Model 1 across loss functions and for the two stocks. Model 4 is a competitor to Model 1 indicating that in some cases omitting trading intensity may not be very detrimental to the performance of the model. It is interesting to observe that trading intensity alone (Model 6) is far from being an optimal specification: it is the interaction of the three regressors that are most helpful to estimate the conditional means of the extreme bounds of the interval. For most of the cases, Model 3, with regressors estimated by the "point value approach", is dominated by Model 1, whose regressors depend on the features of the interval, i.e. centers and range.

[Table 4]

In Table 5 we report Diebold-Mariano tests to formally test the superiority of Model 1 *versus* the parametric linear model and the remaining five non-parametric models for BAC and AMZN at both 5-minute and 1-minute frequencies. Similar tables (S17 – S21) for the remaining five stocks are in the supplementary file. For BAC at both frequencies, there is overwhelming evidence that Model 1 is superior across loss functions; all p-values but a few are practically zero so that we reject the null hypothesis of equally predictive accuracy (in-sample). For AMZN, the evidence is mixed.

At the 5-minute frequency, the non-parametric Models 1, 2, 3 and 4 seem to be equivalent. Model 1 is marginally superior to the linear IAR-TS but superior to Models 5 and 6. At the 1-minute frequency, Model 1 seems equivalent to Model 4 but superior to Models 2, 3, 5, and 6. The evidence with respect to the linear model IAR-TS is mixed and depends on the loss function. When the norm of the loss function is $p = 1$, Model 1 is superior to the linear specification. For the remaining five stocks, we find similar patterns. For the banking stocks, Model 1 outperforms other parametric and nonparametric models in most cases at both frequencies. For the technology stocks, at the 5-minute frequency, Model 1 is still one of the preferred specifications, but at the 1-minute frequency, the linear specification is a contender to Model 1 for GOOG and AAPL stocks.

Overall, Model 1 is a superior specification for the banking stocks at both frequencies and for the technology stocks at the 5-minute frequency; these results confirm that the joint inclusion of the three regressors, i.e. center, range, and intensity, are needed to produce the best fit for the bounds of the interval. For the technology stocks at the 1-minute frequency, we observe that the linear specification and Model 1 seem to be equivalent under quadratic loss functions.

[Table 5]

In Figures 3 – 6, we plot the estimated conditional surfaces of the lowest and highest returns for BAC and AMZN provided by the non-parametric Model 1. Similar figures (Figures S3 – S12) for the remaining five stocks are in the supplementary file. We plot the expected y_{lt} and y_{ut} as a function of $\widehat{r}_t$ and $\widehat{\lambda}_t$ keeping $\widehat{c}_t$ fixed at its sample median. The variable $\widehat{r}_t$ is the conditional expected range and proxies the volatility of the underlying latent return process and the variable $\widehat{\lambda}_t$ is the conditional expected trading intensity. The surfaces clearly indicate the nonlinear behavior of the function. The direction of the arrows in the horizontal axis (volatility and intensity) and in the vertical axis (low and high returns) indicate that the values go from low to high. In general, we find that the relationship of extreme returns with trading intensity and volatility goes in the expected direction. For both frequencies, the higher the volatility and the trading intensity are, the larger the magnitude of the extreme returns are, i.e. the high return goes up and the low return goes down (by examining the surfaces along the diagonals). The response of extreme returns to trading intensity depends on the level of volatility. For low levels of volatility, extreme returns tend to be flat as trading activity intensifies but when the volatility is moderate to high, both extreme returns are very responsive to increasing trading intensity and more so the low returns.

[Figures 3 – 6]

5 Conclusion

In contrast to existing regression-type models for interval-valued data, we have exploited the extreme nature of the lower and upper bounds of intervals to propose a semiparametric model for interval-valued time series data that is rooted in the limiting results provided by the extreme value theory. We have assumed that there are two stochastic processes that generated the interval-valued data. The first process $\{Y_t\}$ is latent, e.g. the process of financial returns, and follows some unknown conditional density. The second process $\{N_t\}$ is observable and it consists of a collection of random draws, e.g. the process of number of trades. In this framework, the upper and lower bounds of the interval, e.g. the highest and the lowest returns at time t, are the realized extreme observations within the sample of random draws at time t. We have shown that the conditional mean of extreme returns is a nonlinear function of the conditional moments of the latent process and of the conditional intensity of the process for the number of draws. This specification provides a natural context to test the relationship between extreme returns and intensity of trading. Asymmetric information models of market microstructure claim that trading volume is a proxy for latent information on the value of a financial asset. With interval-valued time series of 5-minute and 1-minute returns for seven stocks of banks and technology companies, we have found that indeed there is a nonlinear relationship between extreme returns and intensity of trading, which is superior to linear specifications.

The proposed semiparametric model has advantages over the existing models. It is general enough to accommodate linear specifications when these are granted, but the most important advantage is that the model is robust to misspecification of the conditional density of the latent process. We have estimated the conditional mean of the extremes, which is nonlinear on the conditional moments of the latent process, with nonparametric methods. In doing so, we have avoided to choose a specific functional form of the conditional density, which according to extreme value theory is the driver of the nonlinearity. However, the nonparametric function depends on regressors that are generated in a first step. We have shown that the effect of the first-step parameter uncertainty into the second-step nonparametric estimator is asymptotically negligible, and therefore, our two-step estimator has typical nonparametric convergence rate and it is asymptotically normal.

Appendix

A.1 Relaxing the i.i.d. assumption of the random draws $\{N_t\}$

According to extreme value theory for stationary process, the i.i.d. assumption of the random draws within each sampling time interval can be substantially relaxed to strictly stationarity with certain regularity conditions, which allows the y_{it} sequence in $\mathcal{S}_t$ to be weakly dependent without gravely affecting our model specifications. For notational simplicity, we drop the time subscript t. Let $\{Y_i\}$ be a strictly stationary process, where the subscript $i = 1, 2, \ldots, N$ denotes the time order of total N transactions happening during each time period. Let $M_N = \max\{Y_1, \cdots, Y_N\}$ be its sample maximum. Define $\{\widetilde{Y}_i\}$ as the i.i.d. process associated with $\{Y_i\}$ if the two processes share a common marginal distribution function $F(y) = P(Y \leq y) = P(\widetilde{Y} \leq y)$, and $\widetilde{M}_N = \max\{\widetilde{Y}_1, \cdots, \widetilde{Y}_N\}$ as its sample maximum. If the distribution function $F(\cdot) \in \mathrm{MDA}(H)$, the limit distribution of the sample maximum $\widetilde{M}_N$ of the associated iid process $\{\widetilde{Y}_i\}$ is H, i.e., as $N \to \infty$,

$$c_N^{-1}(\widetilde{M}_N - d_N) \xrightarrow{d} H.$$

If the process $\{Y_i\}$ satisfies the conditions $D(u_N)$ and $D'(u_N)$ given below, the limit distribution of the sample maximum M_N of the stationary process $\{Y_i\}$ is also H, i.e., as $N \to \infty$,

$$c_N^{-1}(M_N - d_N) \xrightarrow{d} H,$$

with the same centering and normalizing terms c_N and d_N.

Condition. *$D(u_N)$: for each $y \in \mathbb{R}$ the sequence $u_N = c_N y + d_N$ satisfies that, for any integers p and q, with $p + q$ different numbers picked out from the sequence of the time ordered subscript $i = 1, 2, \ldots, N$ such that $1 \leq i_1 < \cdots < i_p < j_1 < \cdots < j_q \leq N$ and $j_1 - i_p \geq l$, we have*

$$\left| P\left(\max_{i \in A_1 \cup A_2} Y_i \leq u_N\right) - P\left(\max_{i \in A_1} Y_i \leq u_N\right) P\left(\max_{i \in A_2} Y_i \leq u_N\right) \right| \leq \alpha_{N,l},$$

where $A_1 = \{i_1, \cdots, i_p\}$, $A_2 = \{j_1, \cdots, j_q\}$, $\alpha_{N,l} \to 0$ as $N \to \infty$ for some sequence $l = l_N = o(N)$.

Condition. $D'(u_N)$: *for each* $y \in \mathbb{R}$ *the sequence* $u_N = c_N y + d_N$ *satisfies that*

$$\lim_{k\to\infty} \limsup_{N\to\infty} n \sum_{j=2}^{[n/k]} P(Y_1 > u_N, Y_j > u_N) = 0.$$

Condition $D(u_N)$ describes a specific type of asymptotic independence. As a distributional mixing condition, it is weaker than most of the classical forms of dependence restrictions. Condition $D'(u_N)$ means that joint exceedance of U_N by every pairs (Y_i, Y_j) is very unlikely as N approaches to ∞. These two conditions are discussed extensively in Leadbetter, Lindgren, and Rootzén (1983). Direct verification of these two conditions is tedious. However, for Gaussian stationary linear process, $Y_i = \sum_{j=-\infty}^{\infty} \psi_j Z_{i-j}, \quad i \in \mathbb{Z}$, where $\{Z_i\}$ is an iid Gaussian innovation process, the conditions $D(u_N)$ and $D'(u_N)$ boil down to a very weak and intuitive one: the auto-covariance function $\gamma(h) = cov(Y_i, Y_{i+h})$ of the process $\{Y_i\}$ approaches to 0 faster than $(\ln h)^{-1}$ as $h \to \infty$, i.e., $\gamma(h) \ln h \to 0$. It even includes Gaussian fractional ARIMA process with the order of difference $d \in (0, 0.5)$ whose auto-covariances are not absolutely summable. Moreover, the Gaussian distribution of innovations Z_i can be further relaxed to sub-exponential distributions, and the sample maxima of sub-exponential linear process may still have a non-degenerate limit distribution. See Leadbetter and Rootzén (1988) for more details.

A.2 Procedure for cleaning the TAQ database and treatment of outliers

We use the Steps C1 – C4 described below to clean the TAQ high frequency data:

- C1. Keep entries labeled as *regular trade* (which is not corrected, changed or signified as cancel or error with corrected trades) or *corrected trade* (which contains the original time and the corrected data for the trade). They are trades with *Trade.Correction.Indicator* = 00 or 01.

- C2. Delete entries with abnormal Sale Condition. Since the stocks of three banks are CTA issues and the stocks of the four technology companies are UTP issues, their codes for sale condition are slightly different. For the banks, we drop the trades where *Sale.Condition* has a letter code except for '@', F, I, 'Q', 'O', 'M', and '6'. For the technology companies, we drop the trades where *Sale.Condition* has a letter code except for '@', F, I, 'Q', 'OX', 'M', and '6X'. See the Daily TAQ Client Specification (Version 2.2a) for details about sale conditions.

- C3. Let p_{it} be the price of the i-th transaction in the t-th time interval (1 or 5 minutes), and $Q_t(p)$ be the p-th sample quantile of the prices $\{p_{it}\}_{i=1}^{N_t}$ in time interval t, where N_t is the total number of trades in time interval t. Let the Quartile Range (QR) of the price in interval t be $\mathrm{QR}(t; a) \equiv Q_{1-a/2}(t) - Q_{a/2}(t)$ with $a \in [0, 1]$. Set $a_1 = 0.5$ and follow the outlier detection procedure recursively: if $\mathrm{QR}(t; a_n) > 0$, all transactions with trade prices outside of the region $[Q_{a_n/2}(t) - 3/2^{n-1} \cdot \mathrm{QR}(t; a_n), Q_{1-a_n/2}(t) + 3/2^{n-1} \cdot \mathrm{QR}(t; a_n)]$ are removed, and the outlier detection stops. Otherwise, if $\mathrm{QR}(t; a_n) = 0$, go to the next step and set $a_{n+1} = a_n/2$.
- C4. Finally, delete the entries outside the official trading time window (9:30am – 4:00pm).

Note that if Step C3 stops at $a_1 = 0.5$, the outlier detection procedure coincides with the Tukey's fences (Tukey, 1977) used to detect extreme outliers. However, when the trade prices are highly clustered in a time interval such that the inter-quartile range is 0, the Tukey's fences method removes too many transactions and results in no price variation within the time interval. However, our modified procedure generalizes Tukey's fences by enlarging the quantile range if the previous quantile range in the procedure is zero. Hence, compared to the Tukey's fences, the outlier detection procedure in Brownless and Gallo (2006) using trimmed sample mean and standard deviation, and that in Barndorff-Nielsen et al. (2009) using sample median and mean absolute deviation from the mean, our procedure is more conservative in defining outlier in the sample.

Before proceeding with the estimation of the nonparametric models, we introduce two pre-treatments of our sample to deal with two practical data problems in nonparametric estimation.

Global outliers. These are triggered by some economic shocks, which cause extremely low or high returns and a very large numbers of trades in some days. This will produce a very large kurtosis as we have seen in the descriptive statistics of Table 1. Because of the local features of the kernel approach, the results of the local linear estimation are sensitive to the global outliers. To alleviate this problem, we censor the sample of the low and high returns and the covariates for each model at the 0.5% percentile on both sides: for each variable, all observations below (above) its 0.5th (99.5th) sample percentile are set to the 0.5th (99.5th) sample percentile. In comparison to the traditional sample trimming method that excludes observations in the top and bottom percentiles, the sample censoring method does not discard any extreme observation. It keeps the contribution of extreme observations in the tails of the distribution, and thus has a lesser effect on estimates of scale.

Heavy tails. High frequency data exhibit heavy tail behavior with just a few observations in the tail domains of the empirical distributions of the variables. Ignoring heavy tails in the data may lead to serious distortions and large biases in the estimators. To alleviate this problem, we apply the logarithm transformation to the regressors that are positive and have only heavy right tails (range, conditional intensity, number of trades, conditional standard deviation), and the two-side logarithm transformation (A.1) to the center returns that have heavy tails on both sides,

$$S(x_i) = \begin{cases} \log(1 + x_i - x_{\text{med}}), & \text{if } x_i > x_{\text{med}}, \\ -\log(1 + x_{\text{med}} - x_i), & \text{otherwise}, \end{cases} \tag{A.1}$$

where x_{med} is the median value of the sample $\{x_i\}_{i=1}^{T}$. Other approaches dealing with heavy tails in kernel-based nonparametric estimation are more complex and require additional technical assumptions (Markovich, 2008). We choose the fixed transformation because of its simplicity.

We use the following algorithm to obtain the nonparametric estimates for the six proposed nonparametric models: First, censor the sample of dependent variables and the regressors at 0.5% percentile on both sides. Second, apply the fixed transformations to the regressors. Third, obtain the estimates of the conditional mean functions of the transformed sample by the local linear kernel smoothing method with a second order Epanechnikov kernel. The optimal bandwidths are selected by minimizing the least squares cross-validation function for each model and $j = l$ and u,

$$CV_j^{LL}(b_1, \ldots, b_q) = \frac{1}{T} \sum_{t=1}^{T} \left(Y_{jt} - \widehat{M}_{j,-t}^{LL}(\mathbf{x}_t^*) \right)^2,$$

where $\widehat{M}_{j,-t}^{LL}(\cdot)$ is the leave-one-out local linear estimator and $\mathbf{x}_t^*$ includes the q transformed regressors of the model. Finally, the local linear estimate of the conditional mean function is $\widehat{M}_{j,opt}^{L}(T(\mathbf{x}_t))$ where the function $T(\cdot)$ transforms each regressors in $\mathbf{x}_t$ appropriately.

References

Andersen, T. (1996). Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility. *Journal of Finance*, 51, pp. 169–204.

Arroyo, J., González-Rivera, G, & Maté, C. (2011). Forecasting with Interval and Histogram Data. Some Financial Applications, in *Handbook of Empirical Economics and Finance*, A. Ullah and D. Giles (eds.). Chapman and Hall, pp. 247–280.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., & Shephard, N. (2009). Realized kernels in practice: trades and quotes. *Econometrics Journal*, 12(3), pp. 1–32.

Brownless, C.T., & Gallo, G.M. (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis*, 51, pp. 2232–2245.

Chung, K. L. (2001). A course in probability theory (3rd edition). Academic Press.

Conrad, C., & Mammen, E. (2008). Nonparametric regression on latent covariates with an application to semiparametric GARCH-in-Mean models. University of Heidelberg, Department of Economics, Discussion Paper No. 473.

Darrat, A.F., Rahman, S., & Zhong, M. (2003). Intraday trading volume and return volatility of the DJIA stocks: a note. *Journal of Banking and Finance*, 27, pp. 2035–2043.

de Haan, L., & Ferreira, A. (2000). *Extreme Value Theory: An Introduction.* Springer, New York.

Domowitz, I., & White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Journal of Econometrics*, 20, pp. 35–58.

Easley, D., & O'Hara, M. (1992). Time and the Process of Security Price Adjustment. *Journal of Finance,* 47, pp. 905–927.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling Extremal Events: for Insurance and Finance*. Springer, New York, NY.

Engle, R.F. (2000). The Econometrics of Ultra-High-Frequency Data. *Econometrica*, 68, pp. 1–22.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 24, No. 2, pp. 180–190). Cambridge University Press.

Fleming, J., & Kirby, C. (2011). Long Memory in Volatility and Trading Volume. *Journal of Banking & Finance*, 35, pp. 1714-1726.

Gnedenko, B. (1943). Sur la distribution limité du terme maximum d'une série aléatoire. *Annals of mathematics*, 44, 423–453.

González-Rivera, G., & Lin, W. (2013). Constrained Regression for Interval-valued Data. *Journal of Business and Economic Statistics*, 31(4), pp. 473–490.

Komunjer, I., & Owyang, M. (2012). Multivariate Forecast Evaluation and Rationality Testing. *The Review of Economics and Statistics*, 94, pp. 1066–1080.

Lamoureux, C., & Lastrapes, W. (1990). Heteroskedasticity in Stock Return Data: Volume Versus GARCH Effect. *Journal of Finance*, 45, pp. 221–229.

Leadbetter M.R., Lindgren G., & Rootzén H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer, Berlin.

Leadbetter, M.R., & Rootzén, H. (1988). Extreme Theory for Stochastic Processes. *Annals of Probability*, 16, pp. 431–478.

Lin, W., & González-Rivera, G. (2016). Interval-valued Time Series Models: Estimation based on Order Statistics. Exploring the Agriculture Marketing Service Data. *Computational Statistics & Data Analysis*, 100, pp. 694–711.

Markovich, N. (2008). Nonparametric analysis of univariate heavy-tailed data: research and practice (Vol. 753). John Wiley & Sons.

Ning, C., & Wirjanto T.S. (2009). Extreme Returnvolume Dependence in East-Asian Stock Markets: A Copula Approach. *Finance Research Letters*, 6, pp. 202-209.

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *Journal of Business*, 53, pp. 61-65.

Sita, B.B., & Westerholm, P.J. (2011). The Role of Trading Intensity Estimating the Implicit Bidask Spread and Determining Transitory Effects. *International Review of Financial Analysis*, 20, pp. 306–310.

Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley.

Table 1: Descriptive Statistics for Bank of America Corp. and Amazon.com Inc *

Statistics	Returns in Low	Returns in High	Returns in Close	Center	Range	# of trades	Returns in Low	Returns in High	Returns in Close	Center	Range	# of trades
	Bank of America Corp. (BAC) at 5-min Freq.						*Bank of America Corp. (BAC) at 1-min Freq.*					
Minimum	-89.65	-6.63	-78.61	-40.58	0.04	298.00	-59.23	-8.43	-55.38	-29.40	0.04	11.00
1st Quartile	-13.09	4.13	-6.86	-4.34	12.67	1049.50	-6.39	2.09	-3.85	-2.14	4.42	168.00
Median	-7.41	8.01	0.00	0.00	17.02	1438.50	-4.02	4.18	0.00	0.00	8.42	279.00
3rd Quartile	-3.97	13.13	7.14	4.35	25.30	2084.00	-1.79	6.40	3.69	2.17	12.58	446.00
Maximum	5.87	122.49	105.69	61.03	122.92	15702.00	7.85	59.32	59.32	29.45	66.86	3789.00
Mean	-10.20	10.03	0.08	-0.08	20.23	1791.91	-4.56	4.69	0.02	0.06	9.26	355.55
Variance	112.81	106.16	212.75	69.79	158.78	1525810.14	25.65	24.64	46.61	15.69	37.81	86358.38
Correlation	0.28			-0.03			0.25			-0.02		
Skewness	-2.43	3.10	0.20	0.12	2.37	3.01	-2.45	2.38	0.01	-0.03	2.19	2.99
Kurtosis	9.34	19.10	5.72	4.94	9.77	16.21	12.02	11.76	6.47	4.11	9.44	16.35
	Amazon.com Inc (AMZN) at 5-min Freq.						*Amazon.com Inc (AMZN) at 1-min Freq.*					
Minimum	-370.61	-30.28	-171.70	-128.66	2.97	175.00	-370.61	-51.90	-132.79	-142.56	0.01	16.00
1st Quartile	-13.46	2.89	-5.88	-4.34	9.94	388.00	-6.30	1.38	-2.81	-2.02	4.95	72.00
Median	-7.03	5.96	-0.12	-0.35	15.00	590.00	-3.48	3.03	0.00	-0.19	7.28	113.00
3rd Quartile	-3.42	11.29	5.20	2.85	24.12	954.50	-1.70	5.77	2.77	1.58	11.48	190.00
Maximum	2.12	113.29	113.29	42.88	483.89	13600.00	6.92	103.93	95.24	47.23	456.11	9954.00
Mean	-10.73	8.92	-0.57	-0.91	19.64	839.49	-4.95	4.41	-0.12	-0.27	9.36	167.41
Variance	228.45	107.47	209.87	77.31	362.60	713745.60	50.07	28.48	47.72	19.73	78.17	43340.97
Correlation	-0.09			-0.36			0.01			-0.27		
Skewness	-10.29	3.34	-1.19	-3.07	10.30	5.27	-19.55	4.00	-1.86	-5.29	17.66	15.12
Kurtosis	202.05	19.17	20.30	38.04	213.74	50.11	870.04	39.44	41.78	153.64	779.16	588.75

* The sample period is 22 trading days in June 2017. Returns are in the unit of 1 basis point, and the sample size is 1716 and 8580 at 5-minute and 1-minute frequency respectively. For BAC, the raw data set contains 3085563 trades in total. After removal of outliers, the total number of trades is 3074919 for the sample at 5-minute frequency, and 3050589 at 1-minute frequency. For AMZN, the raw data set contains 1441127 trades in total. After removal of outliers, the total number of trades is 1440561 for the sample at 5-minute frequency, and 1436377 at 1-minute frequency.

Table 2: Models for center and range series of low/high return interval for BAC and AMZN at 5 minutes frequency

	BAC		AMZN	
Center series: ARMA models				
	coeff.	s.e.	coeff.	s.e.
intercept			-0.9051	(0.2248)
AR(1)	-0.6943	(0.2164)	0.0579	(0.0241)
AR(2)	0.0098	(0.0299)		
AR(3)	0.0106	(0.0252)		
MA(1)	0.7194	(0.2150)		
σ^2	69.84		77.1	
LogLike	-6076.12		-6161.97	
AIC	12162.24		12329.93	
Range Series: CARR models*				
	BAC		AMZN	
	coeff.	s.e.	coeff.	s.e.
ω	0.1918	(0.0499)	0.0567	(0.0157)
α_1	0.1615	(0.0219)	0.2650	(0.0228)
α_2	0.2486	(0.0384)	0.1337	(0.0322)
α_3	0.1512	(0.0251)	-0.1163	(0.0304)
α_4			0.1705	(0.0365)
α_5			0.1776	(0.0207)
α_6			-0.0592	(0.0308)
β_1	-1.1157	(0.1684)	0.0085	(0.0509)
β_2	0.2269	(0.1411)	0.8488	(0.0293)
β_3	0.9194	(0.1005)	-0.8550	(0.0655)
β_4	0.2320	(0.1418)	-0.3783	(0.0254)
β_5			0.7572	(0.0401)
κ	3.7189	(0.1295)	5.6709	(0.2478)
σ^2	0.4743	(0.0643)	1.3530	(0.1251)
LogLike	-833.01		-635.71	
AIC	1686.02		1299.41	
Ljung-Box test on standardized residuals†				
	statistic	p-value	statistic	p-value
Q(50)	38.0931	0.8912	41.4138	0.8011
Q(100)	89.5671	0.7635	78.8183	0.9419
Q(200)	151.2688	0.9958	170.3999	0.9365

* If the two null hypotheses for distributional parameters $\kappa = 1$ and $\sigma^2 = 0$ are true, the Burr distribution reduces to the exponential distribution.

† For the pseudo-Pearson residuals, the mean and standard deviation are 0.001 and 0.9835 for BAC respectively, and -0.0090 and 0.9322 for AMZN respectively.

Table 3: Models for series of the number of trades for BAC and AMZN at 5 minutes frequency

Series of number of trades: Autoregressive Conditional Intensity models												
	BAC						AMZN					
		coeff.	s.e.		coeff.	s.e.		coeff.	s.e.		coeff.	s.e.
	α_1	0.30	(0.03)	β_1	-0.06	(0.01)	α_1	0.50	(0.02)	β_1	-1.52	(0.04)
	α_2	0.13	(0.01)	β_2	0.04	(0.01)	α_2	0.84	(0.05)	β_2	-1.30	(0.02)
	α_3	0.12	(0.01)	β_3	-0.20	(0.01)	α_3	0.88	(0.04)	β_3	-1.16	(0.02)
	α_4	0.12	(0.01)	β_4	-0.02	(0.01)	α_4	0.92	(0.03)	β_4	-0.93	(0.02)
	α_5	0.11	(0.01)	β_5	-0.15	(0.01)	α_5	0.80	(0.02)	β_5	-0.10	(0.01)
	α_6	0.11	(0.01)	β_6	-0.04	(0.01)	α_6	0.40	(0.02)	β_6	0.10	(0.03)
	α_7	0.11	(0.01)	β_7	-0.30	(0.00)	α_7	0.17	(0.03)	β_7	0.70	(0.01)
	α_8	0.15	(0.01)	β_8	-0.06	(0.01)	α_8	-0.16	(0.04)	β_8	0.95	(0.03)
	α_9	0.10	(0.01)	β_9	-0.31	(0.01)	α_9	-0.30	(0.05)	β_9	0.30	(0.02)
	α_{10}	0.22	(0.01)	β_{10}	-0.66	(0.01)	α_{10}	-0.13	(0.04)	β_{10}	-0.15	(0.01)
	α_{11}	0.25	(0.01)	β_{11}	0.56	(0.01)						
	α_{12}	-0.02	(0.03)	β_{12}	0.15	(0.01)						
				β_{13}	0.10	(0.01)						
	ω	0.07	(0.00)	$1/d$	0.07	(0.00)	ω	0.16	(0.01)	$1/d$	0.08	(0.00)
LogLik.			-12762.11						-11364.11			
AIC			25578.22						22772.22			

Ljung-Box test on standardized residuals†				
	BAC statistic	BAC p-value	AMZN statistic	AMZN p-value
Q(50)	38.0931	0.8912	41.4138	0.8011
Q(100)	89.5671	0.7635	78.8183	0.9419
Q(200)	151.2688	0.9958	170.3999	0.9365

† For the pseudo-Pearson residuals, the mean and standard deviation are 0.0054 and 1.0433 for BAC respectively, and 0.0017 and 1.1379 for AMZN respectively.

Table 4: In-sample model evaluation for 5-minute and 1-minute low/high stock returns

Models	*MSE*		*MLF*		*MDE*		*MSE*		*MLF*		*MDE*	
	Lower	*Upper*	$p = 1$	$p = 2$	$q = 1$	$q = 2$	*Lower*	*Upper*	$p = 1$	$p = 2$	$q = 1$	$q = 2$
	Bank of America Corp (BAC) at 5-min Freq.						*Bank of America Corp (BAC) at 1-min Freq.*					
IAR-TS	96.18	94.08	26.76	379.46	6.69	95.13	21.58	21.12	13.02	85.14	3.25	21.35
IAR-MTS	96.18	95.02	26.78	381.35	6.70	95.60	21.58	21.12	13.02	85.14	3.25	21.35
Model 1	**83.81**	**82.71**	**25.98**	**332.05**	**6.50**	**83.26**	**20.51**	**20.15**	**12.83**	**81.07**	**3.21**	**20.33**
Model 2	92.54	**81.87**	26.55	347.79	6.64	87.20	22.05	21.82	13.27	87.46	3.32	21.93
Model 3	**88.12**	87.05	26.41	349.18	6.60	87.58	21.90	21.15	13.11	85.84	3.28	21.52
Model 4	88.28	82.87	**26.23**	**341.27**	**6.56**	**85.57**	**20.77**	**20.22**	**12.88**	**81.73**	**3.22**	**20.49**
Model 5	98.43	94.86	27.20	385.46	6.80	96.64	22.39	22.21	13.29	88.93	3.32	22.30
Model 6	95.42	93.94	27.29	377.47	6.82	94.68	22.70	21.92	13.37	88.97	3.34	22.31
	Amazon.com Inc (AMZN) at 5-min Freq.						*Amazon.com Inc (AMZN) at 1-min Freq.*					
IAR-TS	184.92	76.98	25.07	522.27	6.27	130.95	40.26	**19.23**	11.83	118.52	2.96	29.74
IAR-MTS	198.55	81.30	26.08	558.12	6.52	139.92	41.20	19.43	12.00	120.80	3.00	30.32
Model 1	170.80	75.10	**24.40**	490.21	**6.10**	122.95	**39.82**	**19.32**	**11.70**	**117.84**	**2.92**	**29.57**
Model 2	**161.33**	75.97	24.44	**473.19**	6.11	**118.65**	40.66	20.55	11.97	121.97	2.99	30.61
Model 3	174.58	**70.48**	24.42	488.48	6.11	122.53	41.37	19.44	11.83	121.16	2.96	30.40
Model 4	**166.60**	75.63	**24.27**	**482.97**	**6.07**	**121.11**	**39.75**	19.56	**11.70**	**118.19**	**2.92**	**29.66**
Model 5	187.08	**75.05**	25.12	522.67	6.28	131.06	42.47	20.76	12.07	125.99	3.02	31.62
Model 6	171.42	82.16	25.30	505.63	6.32	126.79	42.16	22.12	12.46	128.12	3.12	32.14

The numbers in boldface correspond to the two lowest values for each loss function.

Table 5: In-sample Diebold-Mariano tests of Model 1 v.s. other models for BAC and AMZN

		BAC				AMZN			
		5 min. freq.		1 min. freq.		5 min. freq.		1 min. freq.	
		stat.	p-value†	stat.	p-value†	stat.	p-value†	stat.	p-value†
M. 1 vs. IAR-TS	MSE: lower	-3.4963	0.0002	-4.9617	0.0000	-1.4462	0.0741	-0.2056	0.4186
	MSE: upper	-2.8823	0.0020	-3.9187	0.0000	-0.7816	0.2172	0.1141	0.5454
	MLF: $p = 1$	-6.5061	0.0000	-8.3670	0.0000	-3.4584	0.0003	-3.5049	0.0002
	MLF: $p = 2$	-5.3019	0.0000	-7.2969	0.0000	-1.4653	0.0714	-0.1718	0.4318
	MDE: $q = 1$	-6.5061	0.0000	-8.3670	0.0000	-3.4584	0.0003	-3.5049	0.0002
	MDE: $q = 2$	-5.3108	0.0000	-7.2952	0.0000	-1.4609	0.0720	-0.1717	0.4318
M. 1 vs. M. 2	MSE: lower	-4.4332	0.0000	-7.3835	0.0000	1.5531	0.9398	-2.5297	0.0057
	MSE: upper	0.4590	0.6769	-7.6964	0.0000	-0.6909	0.2448	-4.2655	0.0000
	MLF: $p = 1$	-4.4403	0.0000	-14.2372	0.0000	-0.3406	0.3667	-9.2217	0.0000
	MLF: $p = 2$	-3.4633	0.0003	-13.0307	0.0000	1.2993	0.9031	-5.5705	0.0000
	MDE: $q = 1$	-4.4403	0.0000	-14.2372	0.0000	-0.3406	0.3667	-9.2217	0.0000
	MDE: $q = 2$	-3.4595	0.0003	-13.0180	0.0000	1.3069	0.9044	-5.5604	0.0000
M. 1 vs. M. 3	MSE: lower	-2.1999	0.0139	-6.4415	0.0000	-1.7562	0.0395	-4.9706	0.0000
	MSE: upper	-2.2484	0.0123	-5.2470	0.0000	1.9103	0.9720	-0.4560	0.3242
	MLF: $p = 1$	-3.3821	0.0004	-11.9744	0.0000	-0.1522	0.4395	-6.0294	0.0000
	MLF: $p = 2$	-3.7366	0.0001	-9.1912	0.0000	0.2696	0.6063	-5.5925	0.0000
	MDE: $q = 1$	-3.3821	0.0004	-11.9744	0.0000	-0.1522	0.4395	-6.0294	0.0000
	MDE: $q = 2$	-3.7532	0.0001	-9.1821	0.0000	0.2616	0.6032	-5.6478	0.0000
M. 1 vs. M. 4	MSE: lower	-3.1536	0.0008	-4.2970	0.0000	1.7115	0.9565	0.9341	0.8249
	MSE: upper	-0.7667	0.2216	-2.4898	0.0064	-1.8555	0.0318	-3.0974	0.0010
	MLF: $p = 1$	-3.9140	0.0000	-6.1857	0.0000	1.5936	0.9445	0.2110	0.5836
	MLF: $p = 2$	-3.2291	0.0006	-5.3529	0.0000	1.4953	0.9326	-1.8320	0.0335
	MDE: $q = 1$	-3.9140	0.0000	-6.1857	0.0000	1.5936	0.9445	0.2110	0.5836
	MDE: $q = 2$	-3.2414	0.0006	-5.3504	0.0000	1.5096	0.9344	-1.7748	0.0380
M. 1 vs. M. 5	MSE: lower	-4.1230	0.0000	-7.5117	0.0000	-2.9851	0.0014	-4.9595	0.0000
	MSE: upper	-2.9955	0.0014	-7.4763	0.0000	-0.0318	0.4873	-3.9850	0.0000
	MLF: $p = 1$	-8.8585	0.0000	-14.4863	0.0000	-4.8912	0.0000	-12.5930	0.0000
	MLF: $p = 2$	-5.8580	0.0000	-12.7531	0.0000	-3.0183	0.0013	-7.3010	0.0000
	MDE: $q = 1$	-8.8585	0.0000	-14.4863	0.0000	-4.8912	0.0000	-12.5930	0.0000
	MDE: $q = 2$	-5.8684	0.0000	-12.7492	0.0000	-3.0141	0.0013	-7.2860	0.0000
M. 1 vs. M. 6	MSE: lower	-4.6496	0.0000	-9.0240	0.0000	-0.1630	0.4353	-9.3610	0.0000
	MSE: upper	-4.4626	0.0000	-8.2320	0.0000	-4.5134	0.0000	-8.9482	0.0000
	MLF: $p = 1$	-8.0005	0.0000	-16.7451	0.0000	-6.1236	0.0000	-21.4776	0.0000
	MLF: $p = 2$	-7.3554	0.0000	-14.2630	0.0000	-2.2458	0.0124	-13.5271	0.0000
	MDE: $q = 1$	-8.0005	0.0000	-16.7451	0.0000	-6.1236	0.0000	-21.4776	0.0000
	MDE: $q = 2$	-7.3434	0.0000	-14.2588	0.0000	-2.2105	0.0135	-13.7240	0.0000

† The p-values are calculated under the alternative hypothesis $H_a : L_p(e) < L_c(e)$, i.e., our proposed Model 1 has higher predicative accuracy than the other competing models.

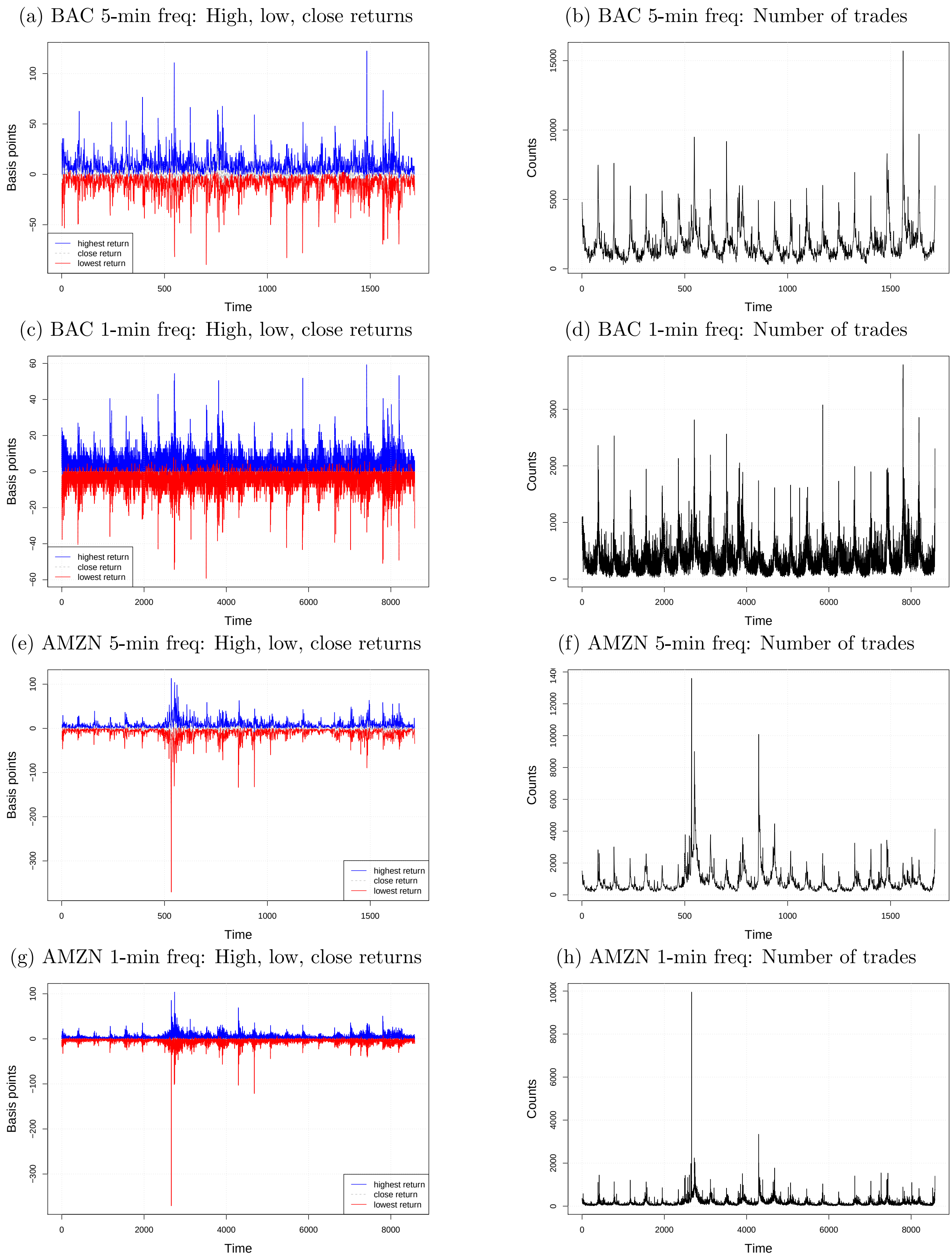


Figure 1: Time series plots for BAC and AMZN at 5-minute and 1-minute frequencies

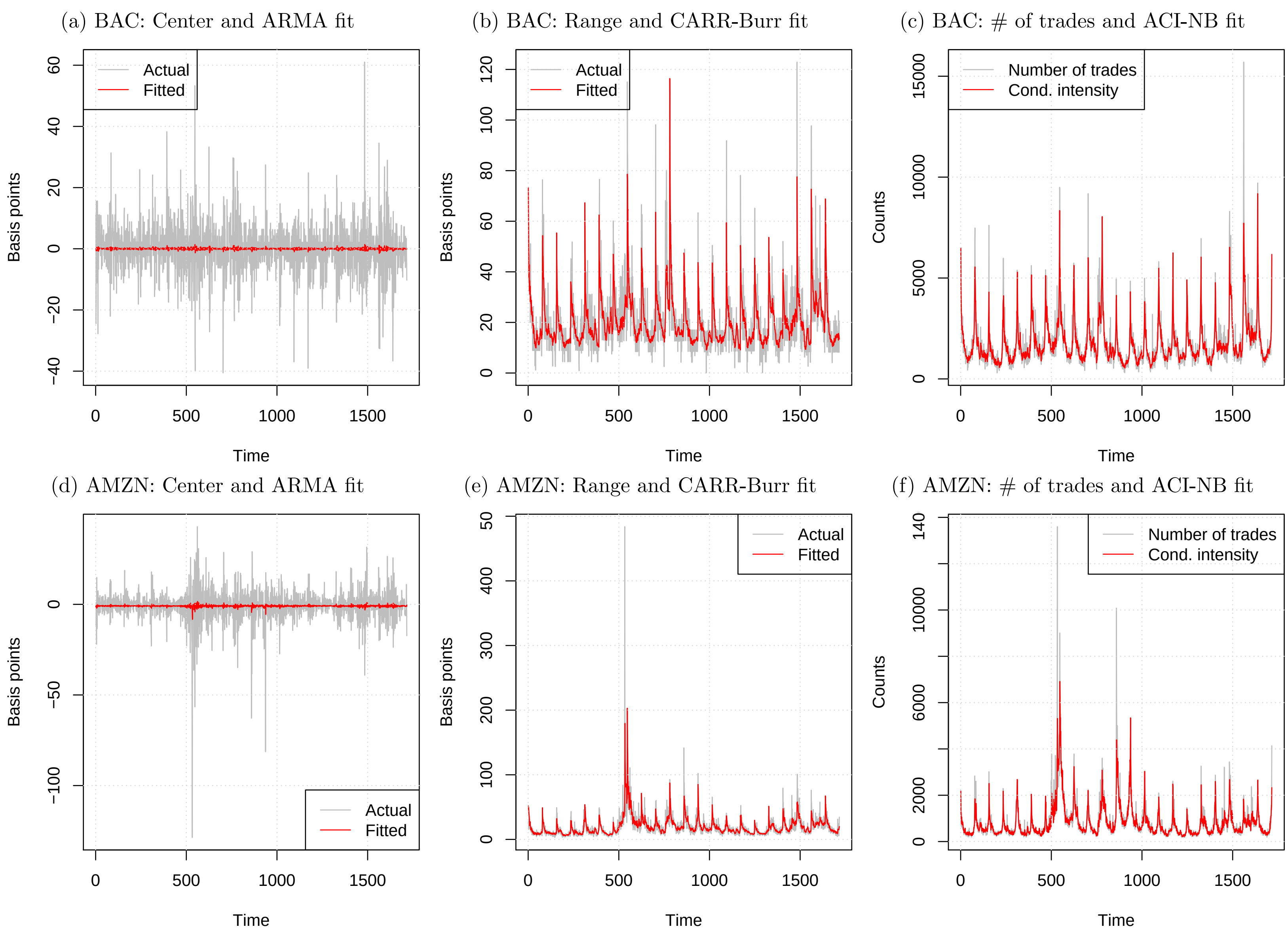
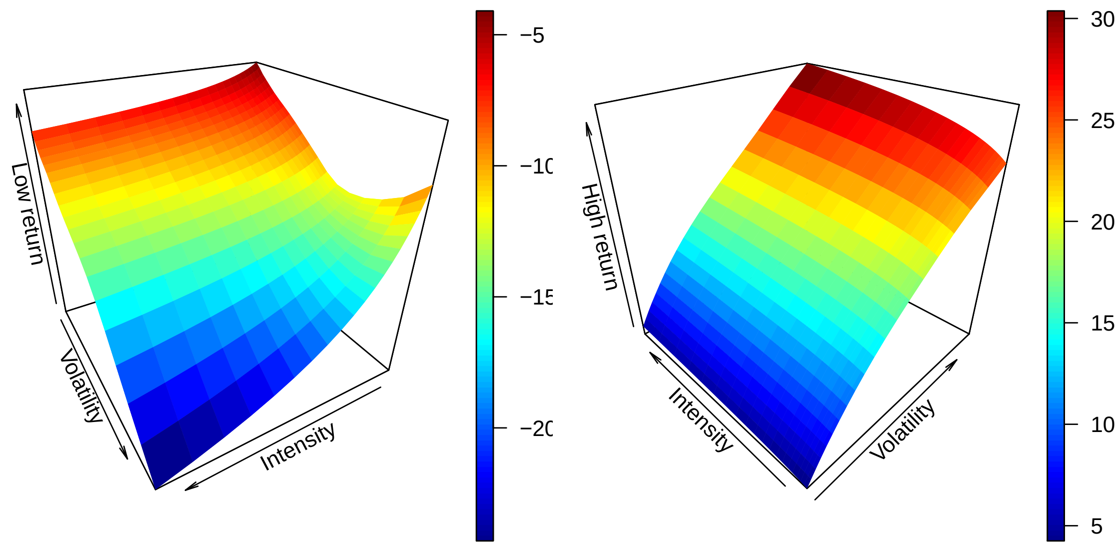


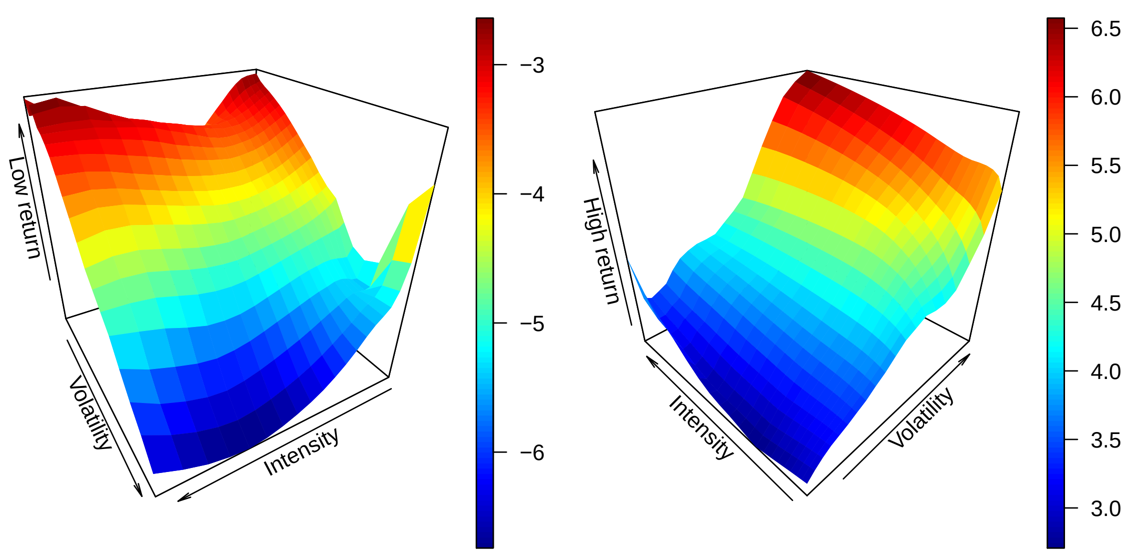
Figure 2: First step estimation for BAC and AMZN at 5-minute frequency



(a) 5-min low return

(b) 5-min high return

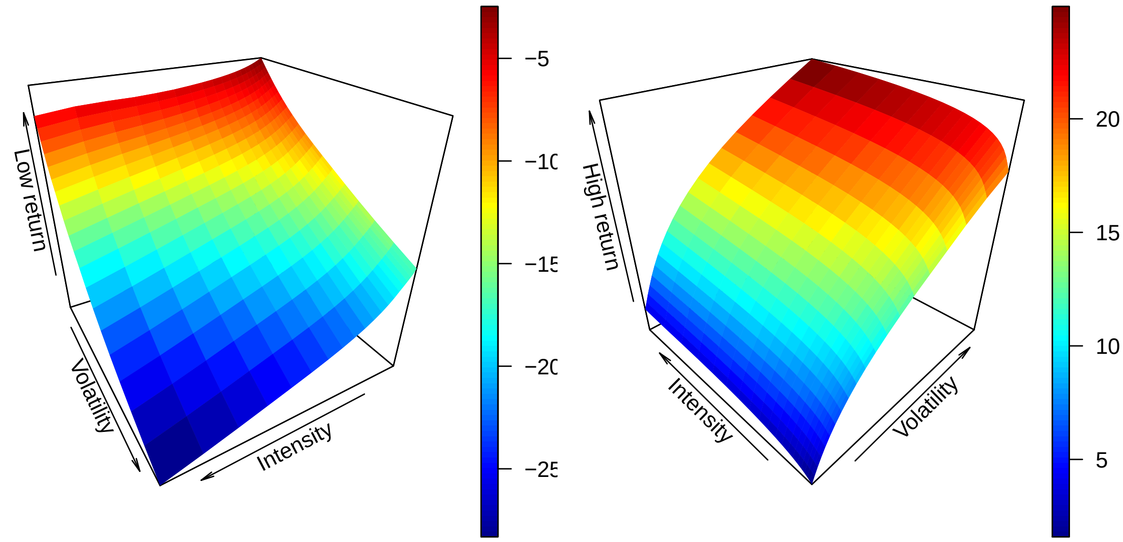
Figure 3: BAC: 5-min low/high returns v.s. intensity and volatility



(a) 1-min low return

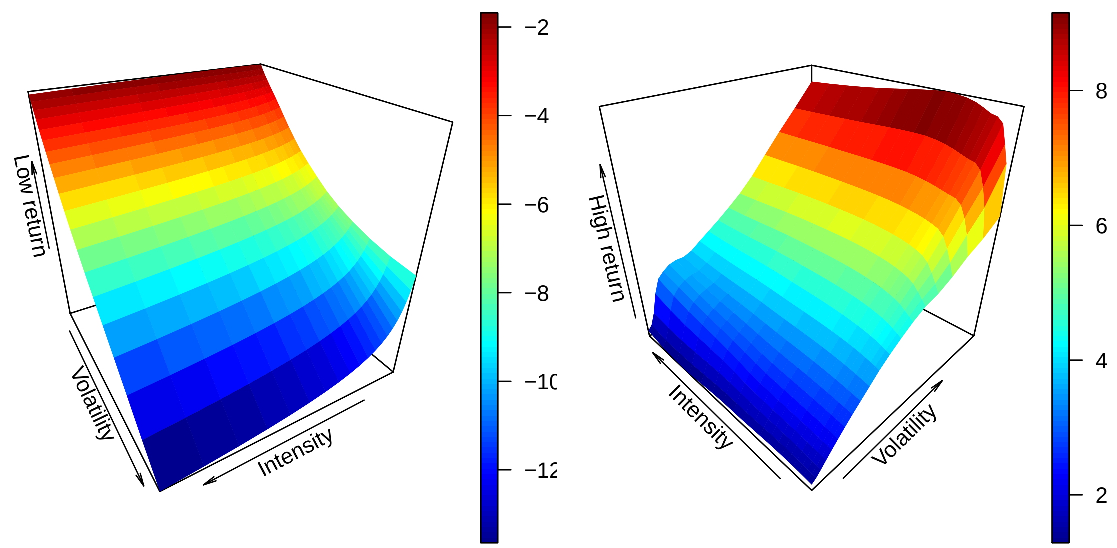
(b) 1-min high return

Figure 4: BAC: 1-min low/high returns v.s. intensity and volatility



(a) 5-min low return (b) 5-min high return

Figure 5: AMZN: 5-min low/high returns v.s. intensity and volatility



(a) 1-min low return (b) 1-min high return

Figure 6: AMZN: 1-min low/high returns v.s. intensity and volatility