MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

„Estimating Jump Variation under Noise with High-Frequency Data“

verfasst von / submitted by

André Thea

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 920
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Quantitative Economics, Management and Finance
Betreut von / Supervisor:	Univ.-Prof. Dr. Nikolaus Hautsch

Estimating Jump Variation under Noise with High-Frequency Data

*Andre Thea**

Department of Finance

University of Vienna

andre.thea@gmail.com

*Data support from the LOBSTER data centre is gratefully acknowledged

Zusammenfassung

Mit Hilfe der neuesten Vor-Mittelwert-Techniken, um Mikrostruktur-Effekte und Ausreißer zu reinigen, verwenden wir Tick-Level-Daten (aufgezeichnet in Millisekunden Genauigkeit), um die Volatilität zu messen. Wir schätzen das Niveau der Sprungvariation (zusätzlich zu der integrierten Variation) im Preisprozess, um mit den Ergebnissen zu vergleichen, die in Christensen, Oomen and Podolskij (2014a) angegeben sind. Wir finden, dass, wenn Sie sich einer kontinuierlichen Zeit nähern, der Bruchteil der Sprungvariation auf etwa 3% von dem 10%-Niveau abnimmt, der aus dem, was jetzt niedrige Frequenz (5-Minuten) Daten geschätzt wird, wie die meisten der jüngsten Literatur gezeigt haben. Dieses Ergebnis kontrastiert leicht mit dem bisher 1% Niveau der Sprungvariation, gefunden durch Christensen, Oomen and Podolskij (2014a). Wir finden, dass für die meisten liquiden Instrumente die Höhe der Sprungvariation nahe bei ihren Ergebnissen liegt. Der Gesamtunterschied in den Sprungvariationsschätzungen ist höchstwahrscheinlich mit den strengeren Datenreinigungstechniken verbunden, die die Anzahl der gereinigten täglichen Renditen reduzieren.

Schlüsselwörter: Hochfrequenzdaten; Sprungvariation; Vor-Mittelung; Realisierte Variation; Bipower Variation; Mikrostrukturgeräusche.

Abstract

Utilizing the latest pre-averaging techniques to clean away microstructure effects and outliers, we use tick-level data (recorded at millisecond precision) to measure volatility. We estimate the level of the jump variation (in addition to the integrated variation) in the price process to compare with results reported in Christensen, Oomen and Podolskij (2014a). We find that as you approach continuous-time, the fraction of jump variation decreases to around 3% from the 10% level estimated from what is now low frequency (5-minute) data, as has been shown by most of the recent literature. This result contrasts slightly with the previously 1% level of jump variation, found by [Christensen, Oomen and Podolskij (2014a)]. We do find that for the most liquid instruments, the level of jump variation is close to their findings. The overall difference in the jump variation estimates is most likely associated to the more rigorous data cleaning techniques that reduce the number of cleaned, daily returns.

Keywords: High-frequency data; Jump variation; Pre-averaging; Realized variation; Bipower variation; Microstructure noise.

Inhaltsverzeichnis

1 Introduction

Extensive literature has attempted to explain the contribution of jumps to the price process of financial instruments. Press (1967) found that jumps represent 20% of the total return variation using monthly data, while more recently Patton and Sheppard (2013) attributed jump variation to just 15% of the total price variation, using returns sampled at 5-minute intervals. Using order-level data, Christensen, Oomen and Podolskij (2014a) found that contrary to these previous empirical findings, the level of jump variation explains just over 1% of the total price variation, attributing the difference to spuriously identified "burstsöf volatility. The topic of this paper is to investigate the claim that if one looks at granular tick-level"data, jumps in the price process of financial securities do indeed embody a much smaller role than has been found empirically. Using pre-averaging techniques from Christensen, Oomen and Podolskij (2014a), we show that the level of jump variation does indeed account for a significantly reduced amount of jump variation, relative to the previous literature, but higher than that found in the Christensen, Oomen and Podolskij (2014a) study, at 3%.

1.1 Outline of the paper

This paper is structured as follows: Section 1 begins with a primer on realized volatility, which describes the standard realized volatility measure, and why it is unsuitable for high-frequency data. The introduction continues with a review of the recent literature and concludes with our motivation for this paper. This is followed by a section on methodology where the proposed methods are explained in greater detail. Next, Section 3 describes the dataset, spanning just under 8 years, that we use in our analysis. Finally, the results are presented in Section 4, followed up by our concluding remarks in Section 5.

1.2 A Primer on Realized Volatility

Realized volatility (RV) is a model-freeëstimator of volatility. So-called model-freeäs no distributional assumptions are used in its derivation. RV is calculated as the sum of squared returns. Andersen and Bollerslev (1998) showed that it is a relatively accurate measure of volatility. Due to its simplicity and accuracy, realized volatility is a popular measure of volatility, widely used in industry.

Definition 1. *Realized volatility is defined as:* $RV(m) := \sum_{i=t}^{m} r_t^2$
where $r_t = p_t - p_{t-1}$ *and* p_t *for* $t = 0, ..., T$*; are logarithmic prices at equally spaced time intervals.*

Realized volatility is easy to compute and asymptotically unbiased in a noiseless setting.

Consider the estimate of RV on a simple, stochastic volatility model with drift.

Definition 2. *Stochastic volatility model with drift:*

$$dS_t = \mu S_t dt + \sigma_t S_t dW_t$$

where μ is the constant drift (expected return) of the price process S_t, σ_t is the stochastic volatility, and dW_t is a standard Wiener Process, for $t \in [0, ..1]$.

Given a time-series of (logarithmic) stock prices p_i over a day with $n+1$ observations such that $i = 0, 1, ..., n$, one could compute the simple realized variance measure, that is, the sum of intra-day squared returns. Using all of the prices available, we use a sampling frequency of length $\Delta = n^{-1}$. This yields the realized variance estimator:

$$RV(n) = \sum_{i=1}^{n} (p_{\Delta} - p_{(i-1)\Delta})^2 = \sum_{i=1}^{n} r_{i\Delta,n}^2$$

Discretizing the continuous time model into time intervals of Δ, we get $r_{i\Delta,n} = \mu\Delta + \sigma(W_{i\Delta} - W_{(i-1)\Delta})$. Deriving the expected value of our realized volatility estimate gives us:

$$\begin{aligned}
\mathbb{E}[RV(n)] &= \mathbb{E}[\sum_{i=1}^{n} (p_{i\Delta} - p_{(i-1)\Delta})^2] \\
&= \mathbb{E}[\sum_{i=1}^{n} ((\mu i\Delta - \mu(i-1)\Delta) + \sigma(W_{i\Delta} - W_{(i-1)\Delta}))^2] \\
&= \mathbb{E}[\sum_{i=1}^{n} (\Delta\mu + \sigma(W_{i\Delta} - W_{(i-1)\Delta}))^2] \\
&= \sum_{i=1}^{n} \Delta^2\mu^2 + \sum_{j=1}^{n} \Delta\mu\sigma \underbrace{\mathbb{E}[(W_{j\Delta} - W_{(j-1)\Delta})]}_{=0} \\
&\quad + \sum_{k=1}^{n} \underbrace{\mathbb{E}[(W_{k\Delta} - W_{(k-1)\Delta})^2]}_{=\Delta} \\
&= \underbrace{n\Delta^2}_{=\Delta}\mu^2 + \underbrace{n\Delta}_{=1}\sigma^2 \\
&= \Delta\mu^2 + \sigma^2 \\
&= \frac{\mu^2}{n} + \sigma^2 \xrightarrow{n\to\infty} \sigma^2
\end{aligned}$$

Hence, asymptotically, i.e., as our sampling frequency tends to infinity, the realized volatility estimator is an unbiased estimator of the true volatility level. In reality, we do not see the continuous price process underpinning our model. The best we can do is to take the trade-by-trade prices. For very high sampling frequencies, i.e., at the order-level, the first term $\frac{\mu^2}{n}$ is negligible, yielding an accurate estimate of the variance of stock prices.

Realized volatility, the square-root of realized variance, is an important estimator as it provides an accurate measure of the level of integrated volatility of a stock price. Volatility is used to measure the riskiness of an asset and so it is very useful in asset management. Stock price volatility has several characteristics such as: mean-reversion and sudden spikes in volatility that cluster before slowly dissipating to the mean, which is asymmetric in response to prices rises and falls. Volatility arising from market downturns is much sharper and more persistent. Jumps in the price process could explain some, if not all, of the volatility spikes. The issue with forecasting volatility is that even if we could predict that a market correction, or jump, would occur at some future point in time, we would find it close to impossible to accurately predict exactly *when* the correction would take place. As a consequence, without insider information, jumps are essentially impossible to forecast (unpredictable). However, by separating the volatility corresponding with the continuous (drift and diffusion) part of the price process with that of the discontinuous (jump) component, we should benefit from improved forecasts.

1.3 Literature Review

Most continuous-time models of the price process of a stock are of a geometric Brownian motion with drift, which implies normally distributed returns. The problem is that across all assets, samples of historical log-prices do not fit a normal distribution; the few, large negative returns, the main source of risk, are non-normal in distribution, and more likely follow a power-law distribution. Hence, we need a continuous-time model that incorporates both these leptokurtic and negatively skewed characteristics of asset price returns. One method is to use a jump-diffusion model as first shown by [Merton (1976), Cox and Ross (1976), Jarrow and Rosenfeld (1984)].

Over the past 50 years or so, the preferred continuous-time models for estimating volatility have been Brownian motion plus jump models. If jumps do indeed account for the excess kurtosis and skewness of stock returns, then their effects should be measurable in the data. Hence, estimating the contribution of jumps to the return variation has been of great academic interest. Table 1.1 below, which was adapted from Fact or Fiction: Jumps at ultra-high frequency, by Christensen K., Oomen R. C. A., and Podolskij M., 2014, *Journal of Financial Economics*

114(3) 576–599, shows various empirical estimates of the level of jump variation, conducted over a period of around 55 years. We can see that the estimated level of jump variation (JV) decreases more or less monotonically with an increasing sampling frequency. Overall, JV has been estimated to be between 10%–20%, higher yet if we focus only on index tracker funds or large-cap stocks.

Tab. 1.1: Empirical estimates of the jump variation component

Article	Data	Period	Frequency	Model	Jump variation
Ball and Torous (1983)	47 US large-cap stocks	1975 - 1977	Daily	JD	50%
Bakshi, Cao and Chen (1997)	S&P 500 options	1988 - 1991	Daily	SVJ	18.9%
Bates (2000)	S&P 500 options	1988 - 1993	Weekly	SVJ	30.3%–38.5%
Andersen, Benzoni and Lund (2002)	S&P 500 spot	1953 - 1996	Daily	SVJ	5.5%
Bollerslev and Zhou (2002)	DM/$	1986 - 1996	Five minutes	SVJ	7.4%
Chernov, Gallant et al. (2003)	DJIA spot	1953 - 1999	Daily	SVJ	9.4%
Maheu and McCurdy (2004)	11 US large-cap stocks	1962 - 2001	Daily	GARCH	29%
	DJIA spot	1960 - 2001	Daily	GARCH	16.6%
Barndorff-Nielsen and Shephard (2004a)	DM/$	1986 - 1996	Five minutes	RV	3.1%
Huang and Tauchen (2005)	S&P 500 futures	1982 - 2002	Five minutes	RV	4.4%
	S&P 500 spot	1997 - 2002	Five minutes	RV	7.3%
Barndorff-Nielsen and Shephard (2006)	DM/$, USD/JPY	1986 - 1996	Five to 120 minutes	RV	5%–21.9%
	S&P 500 spot	1953 - 1996	Daily	SVJ	12.7%
Andersen, Bollerslev and Diebold (2007)	DM/$	1986 - 1999	Five minutes	RV	7.2%
	S&P 500 spot	1990 - 2002	Five minutes	RV	14.4%
	US treasury bonds	1990 - 2002	Five minutes	RV	12.6%
Bollerslev, Law and Tauchen (2008)	40 US large-cap stocks	2001 - 2005	17.5 minutes	RV	12%
	equally weighted index	2001 - 2005	17.5 minutes	RV	9%
Jiang and Oomen (2007)	S&P 500 spot	1987 - 1995	Five minutes	SVJ	18.6%–19.5%
Bollerslev, Kretschmer et al. (2009)	S&P 500 futures	1985 - 2004	Five minutes	RV	6.8%
Todorov (2009)	S&P 500 futures	1990 - 2002	Five minutes	RV	15%
Tauchen and Zhou (2011)	S&P 500 spot	1986 - 2005	Five minutes	RV	5.4%
Andersen, Bollerslev and Huang (2011)	S&P 500 futures	1990 - 2005	Five minutes	RV	4.9%
Corsi and Renò (2012)	S&P 500 futures	1990 - 2007	Five minutes	RV	6%
Patton and Sheppard (2013)	S&P 500 ETF	1997 - 2008	Five minutes	RV	15%

Note. Adapted from Fact or Fiction: Jumps at ultra high frequency, by Christensen K., Oomen R. C. A., and Podolskij M., 2014, *Journal of Financial Economics* 114(3) 576–599. Descriptive note. "JD" refers to the jump-diffusion model of Press (1967) and Merton (1976), "SVJ" to the class of stochastic volatility plus jump models, and "GARCH" to the Generalized Auto Regressive Conditional Heteroskedasticity models, and "RV" to the class of non-parametric measures of variation, including (e.g., realized variance).

A possible explanation for the comparably high estimates, is that the sampling frequency used is too low. An intuitive way of thinking about this is by visualizing it with the following example. Consider the case of a financial asset that is traded once per second, if we sample at a 5-minute frequency we are using just 0.3% of the data available, and the more extreme some price movements seem. The most liquid asset in the data set in our study (see Section 3) was traded at an average rate of 7.4 trades per second over the entire period, so in this case we would be utilizing just 0.045% of the available data if we were sampling at a 5-minute frequency.

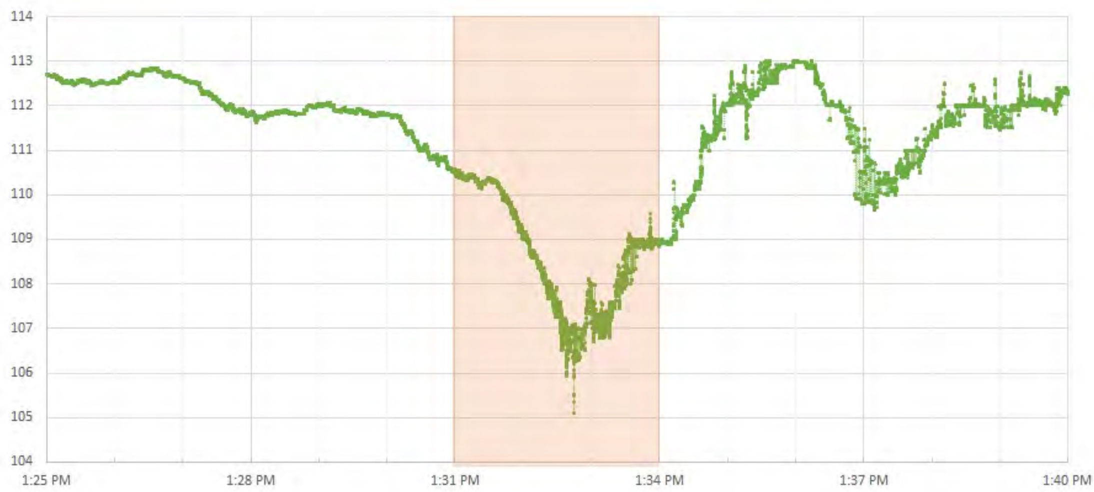
By definition, jumps are instantaneous, discrete price changes. Hence, identifying jumps is, intuitively, best achieved by using the highest sampling frequency data available. This has motivated a shift away from model-based, low-frequency volatility estimators towards non-parametric, high-frequency inference. In an innovative series of papers Barndorff-Nielsen and Shephard (2002b) and Barndorff-Nielsen and Shephard (2004a) introduce the bipower variation estimator as a simple method to estimate the latent integrated variance (continuous component) of the price process. The difference between a realized variance estimate and the bipower variation estimate provides an accurate, yet relatively simple, non-parametric measure of the quadratic variation associated to the jump component. This suggests that as we use data sampled on a finer grid, we should approach the latent variation from above, that is, we should see progressively lower volatility estimates.

Going further, Christensen, Oomen and Podolskij (2014a) hypothesizes that a "burst of volatility", often only seen in real-time, is instead imprecisely identified as a jump, when sampling at the lower frequencies that have been used in previous studies. In Figure 1.1, we can see a macroscopic view of such an extreme event - the flash-crashöf May the 6^{th}, 2010. Notice the huge downward jump in the price of what is an S&P 500 index tracker fund, sampled at a 5-minute frequency (Figure 1.1). This pronounced "jump"becomes a gradual decline when viewed from a chart generated from tick data with hundreds of orders a second (Figure 1.2). This effect is even more pronounced when we rescale the time-axis by a multiple of 5 (Figure 1.3).

Fig. 1.1: The whole day of the "flash"-crash at a 5-minute sampling rate

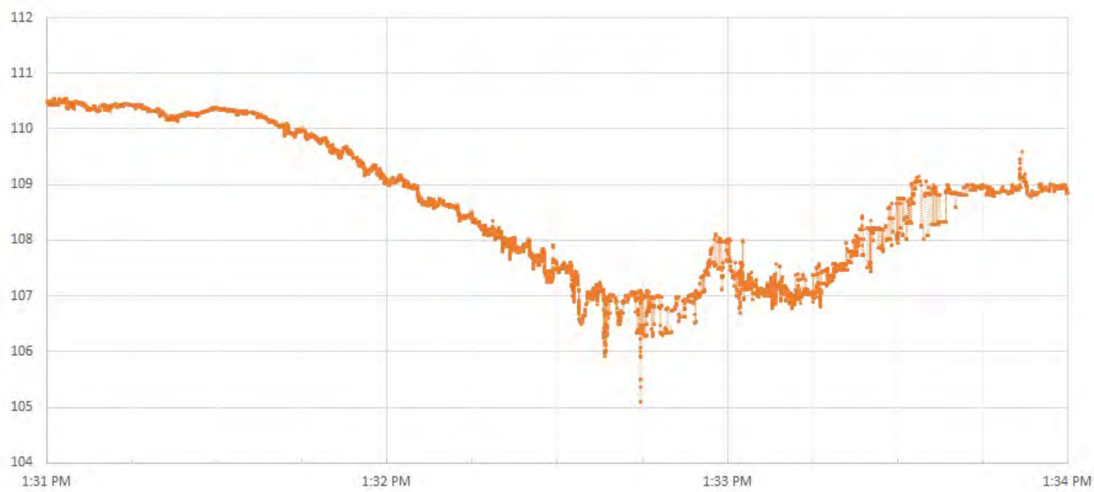
The May 6^{th}, 2010 "flash-crash". A large downward movement, in an S&P 500 index tracker fund, followed by a sharp correction. Note that the sampling times have been adjusted so that the lowest price in the day is shown ($105.11, occurring at 1:32:47pm such that the first price shown is at 9:02:47am). If the prices were shown beginning at 9:00:00am (then 9:05:00am, 9:10:00am and so on), then the lowest price shown would have been $111.81 at 1:30:00pm and we would have missed the full impact of the flash crash. This highlights the importance of using high-frequency data in studying volatility.

Fig. 1.2: Zoom-in on 15 minute surrounding period



The May 6^{th}, 2010 "flash-crash". The abrupt jump, in the S&P 500 index tracker fund jump seen in Figure 1.1(previous figure), is more of a continuous decline when viewing the surrounding 15 minutes with tick-level data (Figure 1.2).

Fig. 1.3: Zoom in on highlighted area over 3-minute period



The price process now looks more like a smooth continuous decline, rather than the abrupt jump we see from the 5-minute horizon. This supports the view of a much reduced role for jumps in the prices process. As there are many more observations with tick-level data than other sampling frequencies, the tick data is able to yield a realistic representation of the continuous price process that is being modeled.

1.4 Research objectives

Against this background, this paper attempts to replicate the findings of Christensen, Oomen and Podolskij (2014a). That is, to estimate the level of jump variation (JV, measured in terms of the level of total variation) in returns of major equity securities. The instruments analyzed are the same as in Christensen, Oomen and Podolskij (2014a), albeit from a different data source. The instruments studied can be split into two main groups: the 30 constituent stocks of the Dow Jones Industrial Average (DJIA), and two index trackers (replicating the S&P 500 index and the NASDAQ-100 index). The results, the level of jump variation, will then be directly compared on both instrument and aggregated levels.

1.5 Motivation

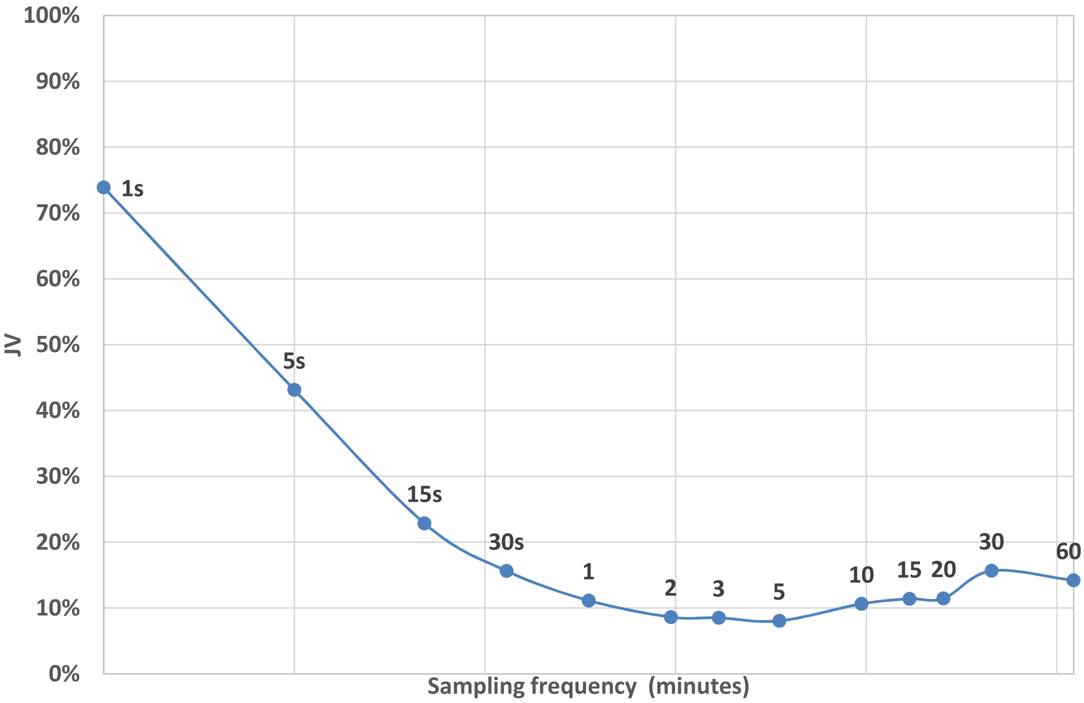
The benchmark has been moved in recent years from what used to be considered high-frequency data (5-minute sampling rate) to the current level (tick-level data). This difference in the sampling frequency could explain, to a certain extent, the discrepancies in the empirical findings. As the sampling frequency of returns, used to derive volatility estimates in empirical studies, has increased, the level of jump variation detected has decreased (see Table 1.1). The scientific consensus has, until only recently, been using 5-minute sampled data in their realized volatility estimates.

The daily volumes of stock trades in world markets has risen substantially over the past decade. Easier access to financial markets from populations in Asia, increased globalization and the rise of high-frequency trading are just a few of the compounding reasons why trade volumes have, and will continue to, rise.

The most liquid stock on the Dow Jones Industrial Average (DJIA) Apple Inc., has averaged 60,000 executed orders daily over an 8-year period. Over a 6-and-a-half hour trading day this corresponds with an average rate of 150 trades a minute. Sampling at a rate of once every 5 minutes would utilize only 0.1% of the available data. This yields an incomplete picture. Recall the example already shown, the May 6^{th}, 2010 flash-crash. When viewed from a macroscopic perspective (a 5-minute sampling frequency, Figure 1.1) the price seems to jump downwards. This "jumplooks more like a steep, but continuous, descent when viewed from a microscopic, trade-by-trade, viewpoint (Figures 1.2 and 1.3). Hence, finding a solution for dealing with the contaminating noise effects should yield much improved estimators of volatility. This paper is evaluating the reliability of a new noise-freeëstimator of volatility.

The obstacle that limited the use of trade-by-trade (tick level) data is the noise, which mostly emanates from market microstructure effects (Hansen and Lunde (2006)). Figure 1.4 shows the contribution of jump variation to the total variation, over increasing sampling frequencies. Note that for sampling frequencies of less than a minute, the noise dominates the signal of the jump component. Using new, pre-averaging techniques (Christensen, Oomen and Podolskij (2014a)) that smooth the data to negate the impact of microstructure effects, such as bid-ask bounce and price discreteness, we are able to fully exploit the available data and get a more accurate image of the latent, integrated volatility.

Fig. 1.4: JV sampling frequency-signature plot for MSFT



Graph displaying the contribution of the jump variation to the total variation (JV) over various sampling frequencies. JV is difference between realized (RV) and bipower variation (BV). RV and BV are the estimators described in Barndorff-Nielsen and Shephard (2004a). The estimates are derived from the period of the 27^{th} of June 2007 through to the 1^{st} of March 2015.

In this regard, the importance of this proposed study cannot be underestimated. If the results are consistent with that of Christensen, Oomen and Podolskij (2014a), then this implies that jumps in the price process of stocks are not the reason for the non-normality of (leptokurtic) log-returns of stock prices. This means that the jump-diffusion models (Merton (1976)) used by practitioners would not fit with the empirical data. Alternative models that capture the leptokurtic behavior of stock returns, should then be a focus for further research.

2 Methodology

This section describes the methodology used in this study. First, a primer on jump diffusion models is presented as an introduction to the topic. Then, a short section on the efficient price process is shown, followed by a third part on the observed price process and how noise contaminates the latent price signal. Building on the framework presented, the last two sections focus on how the jump component is measured on both the latent and, more importantly, on the observed price process.

Certain characteristics of tick data such as drift, stochastic volatility, discontinuous jumps, market microstructure effects (in the form of noise) and outliers should be incorporated into the model. We achieve this as follows: first, we model the *unobserved*, efficient price process as a continuous-time process. Then, we add outliers and the microstructure effects to the latent price process, which combined will equate to the *observed* prices process. Given this framework, we will then infer the magnitude of the jump (and continuous) component(s) from the discretely sampled and noisy measurements of the observed price process.

2.1 A primer on Jump Diffusion Models

Jump diffusion models are constructed as a geometric Brownian motion with drift. This process constitutes two components: the continuous (diffusion) part, and a compound Poisson process modeling the discontinuous (jumps) part.

2.1.1 Poisson process

Definition 3. *The Poisson process is defined as follows:*

Consider a sequence $\{\tau_i\}_{i\geq 1}$ *of independent exponential random variables with mean rate* λ. *Let* $P[\tau_i \geq y] = e^{-\lambda y}$ *be the cumulative distribution function and* $T_n = \sum_{i=1}^{n} \tau_i$. *The process*

$$N_t = \sum_{n\geq 1} 1_{t\geq T_n}$$

is called the Poisson process with parameter λ.

The Poisson process is a càdlàg (right continuous with left limits) process with jump size 1. The Poisson process, alongside the Brownian motion process, is characterized by independent and stationary increments, and hence is a Lévy process.

Definition 4. *A stochastic process* $X = \{X_t : t \geq 0\}$ *is a Lévy process if it satisfies the following properties:*

1. $X_0 = 0$, *almost surely.*
2. *Independent increments: For any* $0 \leq t_1 < t_2 < \cdots < t_n < \infty$, $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$ *are independent.*
3. *Stationary increments: For any* $s < t$, $X_t - X_s$ *is equal in distribution to* X_{t-s}.
4. *Continuity in probability: For any* $\epsilon > 0$ *and* $t \geq 0$ *it holds that* $\lim_{h\to 0} P(|X_{t+h} - X_t| > \epsilon) = 0$.
5. *If* X *is a Lévy process then one may construct a version of* X *such that* $t \mapsto X_t$ *is almost surely right continuous with left limits.*

2.1.2 Compound Poisson process

The *compound* Poisson process is a generalization of the Poisson process. The jump magnitudes have an arbitrary distribution and the waiting times between the jumps are exponential. Essentially this translates to a model where the price process can have varying jump sizes.

Definition 5. *Let* N *be a Poisson process with parameter* λ *and* $\{Y_i\}_{i\geq 1}$ *be a sequence of independent random variables. The compound Poisson process is defined as:*

$$X_t = \sum_{i=1}^{N_t} Y_i$$

Its trajectories remain right-continuous with left limits, but the jump sizes are now random.

2.2 The efficient price process

In this section we look at how the latent (unobserved) price process is modeled. In the model, we assume that the asset price's underlying (time-invariant) continuous process, is defined by $\mathbf{X} = (X_t)_{t\geq 0}$, which is the logarithmic price of an asset that is defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ and is adapted to the filtration $\mathcal{F}_t$ that represents all of the information available up to and including time t. Furthermore, we assume that $\mathbf{X}$ can be decomposed as a local martingale and a càdlàg process. Hence, $\mathbf{X}$ belongs to the class of semimartingales (e.g. Back (1991); Delbaen and Schachermayer (1994)). Combining these properties with the aforementioned stochastic volatility and price jump characteristics, $\mathbf{X}$ can be modeled by the following jump-diffusion stochastic differential equation:

$$dX_t = \mu_t dt + \sigma_t dW_t + (J-1)dN(t) \tag{1}$$

where X_t is the logarithm of the asset price, μ_t is the drift function and σ_t is the adapted càdlàg diffusion function (stochastic volatility), W_t is a standard Brownian (Wiener) process, J is a sequence of non-zero random variables (the jump size), and $N(t)$ is a counting process (the jump intensity).

This can be equivalently represented by:

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + \sum_{i=1}^{N_t^J} J_i \ . \tag{2}$$

The given model is completely non-parametric; apart from the condition of finite variance (a regularity condition needed for the proof of the central limit theorem), we do not put any assumptions on the distributions of the processes. This is incredibly advantageous as it fits with many continuous-time models (e.g. Alizadeh, Brandt and Diebold (2002) with a Ornstein-Uhlenbeck process for log-volatility, Andersen, Benzoni and Lund (2002) for stochastic volatility models with log-normal jumps with a non-homogeneous Poisson process). The model could also be extended to incorporate jumps in the volatility process (e.g. Eraker, Johannes and Polson (2003)).

2.3 The observed price process

We are not able to observe the efficient price process in continuous time, but we do have discreetly sampled measurements available that, at the individual order-level, can contain a substantial amount of noise. The noise is modeled to assist in improving our measurement of the unobserved price process.

One source of noise is related to market microstructure effects that emanate from market imperfections such as bid-ask bounce, price discreetness, latency and informational asymmetries. Bid-ask bounce is an effect that can be observed even when the latent volatility is zero; you still have buyers and sellers trading and consequently you observe prices at bid or ask prices, not at the mid-price, inflating the observed volatility measures. A behavioral example of market microstructure noise can result from prices hitting certain nominal levels, often round-number prices (e.g. multiples of 5), causing them to move suddenly as large number of traders typically have punched in stop-loss orders at these levels. Outliers are another source of noise. These short-lived off-market prices, outside of the National Best Bid and Offer (NBBO), can be attributed to delayed trade reporting on block trades, in addition to fat-finger errors, bugs in the data feed, etc.

With this in mind, and restricting our attention to the unit time interval $t \in [0, 1]$, we model the observed tick data process as:

$$Y_{i/N} = X_{i/N} + u_{i/N} + O_{i/N}, \qquad i = 0, 1, \ldots, N \tag{3}$$

where u is an identically and independently distributed microstructure noise process with $\mathbb{E}(u) = 0$ and $\mathbb{V}(u) = \mathbb{E}(u^2) = \omega^2$, and u is *independent* of X (Diebold and Strasser (2008)). The outlier process is given by $O_{i/N} = \mathbb{1}_{\{i/N \in \mathcal{A}_N\}}\ \mathcal{S}_i$, where $\mathcal{A}_N$ is the set of the appearance events of the outliers, and their magnitudes are given by $(\mathcal{S}_i)_{i=1,\ldots N_1^O}$. We assume that $\mathcal{A}_N$ is $a.s.$ finite and we model it by

$$\mathcal{A}_N = \frac{[NT_i]}{N} |\ 0 \leq T_i \leq 1, \tag{4}$$

where $(T_i)_{i=1,\ldots,N_1^O}$ are the arrival times of another counting process $N^O = (N_t^O)_{t \geq 0}$. Furthermore, we assume that O is mutually independent of $\mathbf{X}$ and u, $O \perp\!\!\!\perp (\mathbf{X}, u)$. This implies that the two counting processes generating jumps (N^J) and outliers (N^O) are mutually independent ($N^J \perp\!\!\!\perp N^O$). Hence, the probability of observing both a jump and outlier simultaneously is negligible.

2.4 Measuring the jump component

Given the framework we have defined above, we can begin to formulate a method for calculating the level of jump variation. Note that we cannot directly estimate the level of jump variation, as it is a latent variable. We must progress by estimating all other components, yielding the jump variation as the remaining component. We begin by defining the quadratic variation, as:

$$[X]_1 = \int_0^1 \sigma_s^2 ds + \sum_{i=1}^{N_1^J} J_i^2. \tag{5}$$

As shown previously, a consistent estimator of the quadratic variation is the realized variance (RV) measure described by Barndorff-Nielsen and Shephard (2004b):

$$RV = \sum_{i=1}^{N} |r_i|^2 \xrightarrow{p} [X]_1, \qquad as \qquad N \to \infty \tag{6}$$

We define JV as the level of quadratic variation associated with the jump component as a fraction of the quadratic price variation as defined below:

Definition 6.

$$JV = \frac{[X]_1 - \int_0^1 \sigma_s^2 ds}{[X]_1} \tag{7}$$

A jump-robust and consistent estimator of the integrated variance, $\int_0^1 \sigma_s^2 ds$, can be estimated using the bipower variation measure (Barndorff-Nielsen and Shephard

(2004b)), given by:

$$BV = \frac{N}{N-1}\sum_{i=2}^{N}|r_{i-1}||r_i| \xrightarrow{p} \int_0^1 \sigma_s^2 ds, \qquad as \qquad N \to \infty \tag{8}$$

Using the estimators yielded from equations 6 and 8, we can isolate the jump variation:

$$RV - BV = \sum_{i=1}^{N_1^J} J_i^2$$

This estimation approach of JV has been used extensively in empirical studies to estimate the jump variation (see Table 1.1 for literature). With lower sampling frequencies, it becomes near impossible to distinguish jumps from variations emanating from the diffusive component. An example of this has already been shown in Figure 1.4. Therefore, in this paper, in order to estimate the jump component, we make use of the highest sampled data that is available.

At the trade-by-trade level, microstructure effects begin to dominate the signal from returns, and thus they cause the standard RV and BV measures described above to be inflated.

The expected value of realized variance sampled at a rate of $\frac{1}{N}$, in a noisy setting is:

$$\begin{aligned}
\mathbb{E}(RV(N)) &= \mathbb{E}[\sum_{i=1}^{N}(Y_{\frac{i}{N}} - Y_{\frac{(i-1)}{N}})^2] \\
&= \mathbb{E}[\sum_{i=1}^{N}(X_{\frac{i}{N}} - X_{\frac{(i-1)}{N}})^2 + 2.(X_{\frac{i}{N}} - X_{\frac{(i-1)}{N}})(u_{\frac{i}{N}} - u_{\frac{(i-1)}{N}}) + (u_{\frac{i}{N}} - u_{\frac{(i-1)}{N}})^2] \\
&= [X]_1 + 2\sum_{i=1}^{N}\mathbb{E}[(X_{\frac{i}{N}} - X_{\frac{(i-1)}{N}})(u_{\frac{i}{N}} - u_{\frac{(i-1)}{N}})] + \sum_{i=1}^{N}\mathbb{E}[(u_{\frac{i}{N}} - u_{\frac{(i-1)}{N}})^2] \\
&= [X]_1 + 2\sum_{i=1}^{N}\underbrace{\mathbb{E}[(X_{\frac{i}{N}} - X_{\frac{(i-1)}{N}})(u_{\frac{i}{N}} - u_{\frac{(i-1)}{N}})]}_{=0} \\
&\quad + \sum_{i=1}^{N}(\underbrace{\mathbb{E}[u^2_{\frac{i}{N}}]}_{=\mathbb{V}[u_{\frac{i}{N}}]} + \underbrace{\mathbb{E}[u^2_{\frac{(i-1)}{N}}]}_{=\mathbb{V}[u_{\frac{(i-1)}{N}}]} - 2\underbrace{\mathbb{E}[u_{\frac{(i-1)}{N}}u_{\frac{i}{N}}]}_{=0}) \\
&= [X]_1 + 2N\omega^2
\end{aligned} \tag{9}$$

Equation 9 shows us that in a noisy diffusion setting, realized variance increases linearly with the sampling frequency. The integrated variation remains constant under the assumption of independently distributed noise process. This leads to a JV estimator that approaches the level of the integrated volatility (see Figure 1.4).

Hence, to estimate $[X]_1$ and its components using tick data, we adopt the pre-averaging approach introduced by Jacod, Li et al. (2009), Podolskij and Vetter (2009a) and Podolskij and Vetter (2009b). Essentially, this method locally smooths the observed price process series Y so that the microstructure component u almost disappears under averaging. The outlier component O is dealt with via the same mechanism, in addition to the data cleaning techniques applied to the raw order book data.

The pre-averaging process produces the smoothed returns, which are used to create consistent estimators of the diffusive- and jump-variation components. These returns $r^*_{i,K}$ are derived from pre-averaged prices in a local neighborhood of K observed prices, i.e.

$$r^*_{i,K} = \frac{1}{K}\left(\sum_{j=K/2}^{K-1} Y_{\frac{i+j}{N}} - \sum_{j=0}^{K/2-1} Y_{\frac{i+j}{N}}\right) \tag{10}$$

where $K \geq 2$ and K is even. In order for the asymptotic distribution of (RV^*, BV^*) to converge to the multivariate normal (see Proposition 2.1), it is necessary that the pre-averaging horizon (θ) grows as the sampling frequency $(\frac{1}{N})$ increases, but at at slower rate. This condition is satisfied by setting $K = \theta\sqrt{N} + o(N^{-\frac{1}{4}})$. We take $K = \lceil\theta\sqrt{N}\rceil$, rounding to the nearest, larger multiple of two. Note that these returns are calculated on overlapping windows of prices of size K. With these pre-averaged returns we are able to construct noise- and outlier-robust estimators of the quadratic and integrated variation as:

$$RV^* = \frac{N}{N-K-2}\frac{1}{K\psi_K}\sum_{i=0}^{N-K+1}|r^*_{i,K}|^2 - \frac{\hat{\omega}^2}{\theta^2\psi_K} \tag{11}$$

$$BV^* = \frac{N}{N-K-2}\frac{1}{K\psi_K}\sum_{i=0}^{N-K+1}|r^*_{i,K}||r^*_{i+K,K}| - \frac{\hat{\omega}^2}{\theta^2\psi_K} \tag{12}$$

where $\psi_K = (1+2K^{-1})/12$. The term $\frac{\hat{\omega}^2}{\theta^2\psi_K}$ corrects for the residual microstructure effects that remain after the pre-averaging process. $\hat{\omega}^2$ is an estimator of the noise variance ω^2. While it can be estimated in a number of ways, we follow Christensen, Oomen and Podolskij (2014a) in selecting the estimator proposed by Oomen (2006), where it is estimated with:

$$\hat{\omega}^2_{AC} = -\frac{1}{N-1}\sum_{i=2}^{N} r^*_i r^*_{i-1}, \tag{13}$$

where $r_i^* = Y_{\frac{i}{N}} - Y_{\frac{i-1}{N}}$.

Proposition 2.1, formulated by Christensen, Oomen and Podolskij (2014a), estimates the magnitude of the jump component in the presence of noise. It gives the joint asymptotic distribution under the null hypothesis of no jumps in the probability limit.

Proposition 2.1. *Assume that the observed price process Y follows Equation 3, and that $\mathbb{E}(u^4) < \infty$. As $N \to \infty$, it holds that*

$$RV^* \xrightarrow{p} [X]_1, \;\; BV^* \xrightarrow{p} \int_0^1 \sigma_s^2 ds. \tag{14}$$

Moreover, suppose that $\mathbb{E}(u^8) < \infty$, and that X is a continuous semimartingale, that is, X follows Equation 5, but with $N_t^J \equiv 0$ for all t. As $N \to \infty$, it holds, furthermore, that

$$N^{1/4} \begin{pmatrix} RV^* - \int_0^1 \sigma_s^2 ds \\ BV^* - \int_0^1 \sigma_s^2 ds \end{pmatrix} \xrightarrow{d_s} MN(0, \Sigma^*), \tag{15}$$

a mixed normal distribution with conditional covariance matrix Σ^* □

The consistency of the pre-averaged estimators RV^* and BV^*, as shown in Equation 14, implies that even in the presence of noise and outliers, we can consistently estimate the magnitude of the jump component:

$$RV^* - BV^* \xrightarrow{p} \sum_{i=1}^{N_1^J} J_i^2.$$

Furthermore, Equation 15 gives the basis for constructing noise-robust, non-parametric jump tests.

2.5 BV^* overestimation in the presence of outliers

In this section, the calibration of the pre-averaging horizon is described. A simulation study was performed in the Christensen, Oomen and Podolskij (2014a) paper in order to check the performance of the noise- and outlier-robust estimators. The pre-averaging horizon, θ, was calibrated using a Monte Carlo study. Different stochastic price processes are modeled (vanilla Brownian Motion, Stochastic Volatility, Brownian Motion with Jumps and with Outliers), with both *i.i.d* and dependent noise processes. The noise process was modeled as the following $AR(1)$ model:

$$\mu_i = \beta \mu_{i-1} + \epsilon_i,$$

where $\epsilon_i \overset{i.i.d}{\sim} N(0, \omega^2(1-\beta^2))$. The memory parameter ($\beta$) is 0.77 in the dependent processes and 0 in the $i.i.d$ simulations. The noise variance, ω^2, is estimated using the noise ratio parameter:

$$\gamma = \sqrt{\frac{N\omega^2}{\int_0^1 \sigma_s^2 ds}},$$

where γ is set as 0.5 and N is set as $40,000$.

Each model was simulated along 10,000 different price paths with N discretized price points. The resulting RV^* estimators remained unbiased in the simulations, but the BV^* estimator was upwardly biased in the Brownian Motion with Outliers and even more so in the Brownian Motion with Jumps model (where it exceeded 10% of the expected value for some pre-averaging horizon parameters, θ). The optimal θ was found to be in the range of 0.5 and 2 [Christensen, Oomen and Podolskij (2014a)].

2.5.1 BV_τ^* estimation algorithm

In this section we describe the method of removing the effect of outliers. To remove the finite sample bias, we implement a threshold filtering method devised in Corsi, Pirino and Renò (2010). The pre-averaged returns that are above a certain threshold in absolute value, $|r_{i,K}^*| > \tau$, are identified; usually it is a sequence of pre-averaged returns that exceed the threshold (see Figure 2.1). We then find and delete the largest noisy return that is used in the estimation of the sequence of pre-averaged returns (the circled return in Figure 2.2), then reconstitute the price series and re-compute the pre-averaged returns. This process is repeated until there are no more threshold violations. We then use these freshly pre-averaged returns to calculate the new bipower variation estimate, BV_τ^*. The threshold is defined as:

$$\tau = \frac{q_\alpha}{N^\varpi}\sqrt{\psi_K \theta \sigma^2 + \frac{\omega^2}{\theta}}, \tag{16}$$

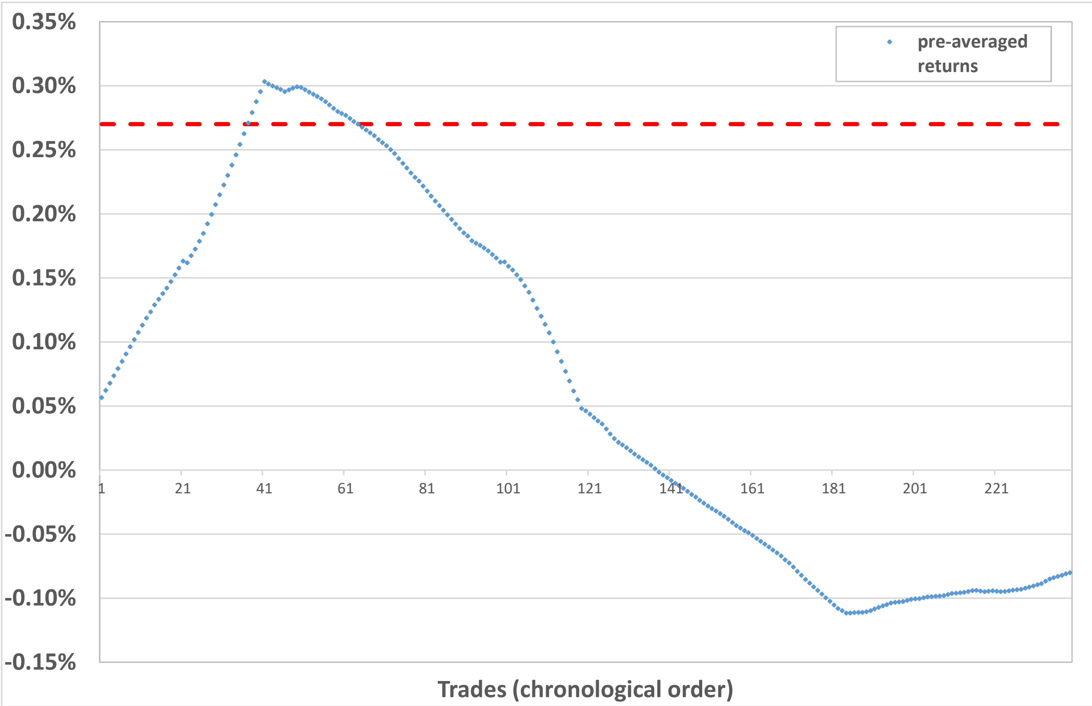
where q_α is the α quantile from the standard normal distribution, and $\varpi \in (0, 0.25)$.

The α used in this paper is set at the 99.9% level, in order to filter out only the most extreme outliers.

Once this process of removing outliers no longer detects any more, BV_τ^* is then calculated using Equation 12 and these pre-averaged returns.

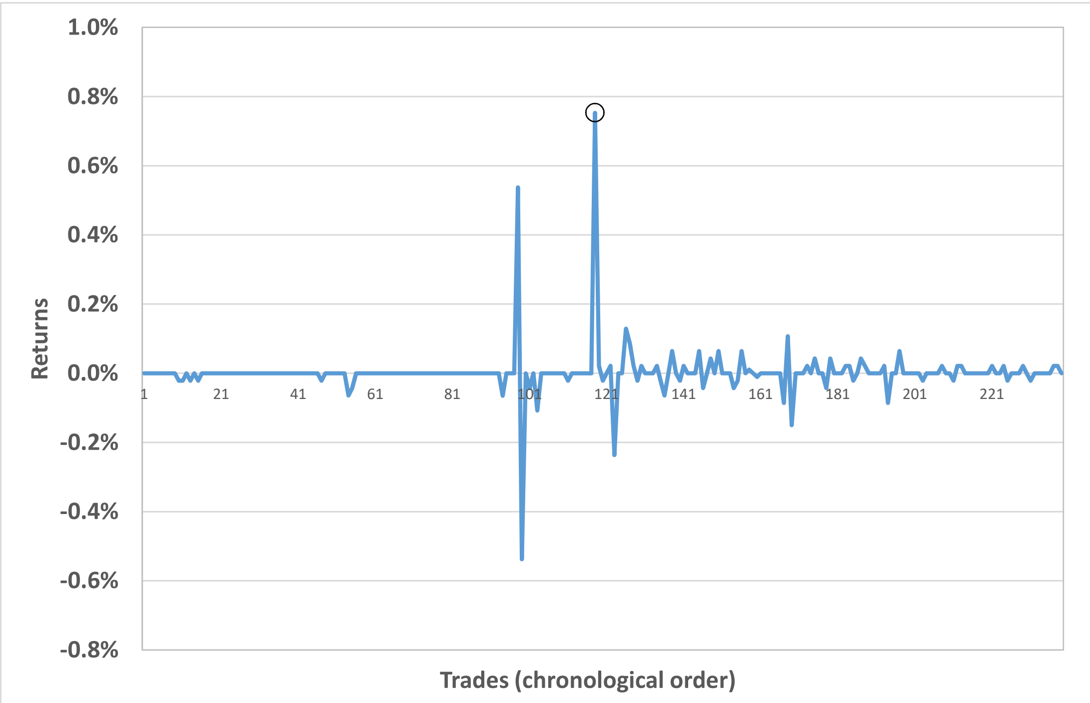
Fig. 2.1: Pre-averaged returns - MSFT 10th September, 2010

.



A chart displaying pre-averaged returns, $r^*_{i,K}$ of Microsoft stock taken from the morning of Thursday, the 10th of September 2014. The dotted red line, symbolizes τ, the threshold for outlying pre-averaged returns. From the set of noisy returns used in the computation of all the pre-averaged returns that are greater than the threshold (trades 37-64), the largest is identified. This largest noisy return is removed and the noisy price time-series is reconstituted from the remaining returns (see Figure 2.2)

Fig. 2.2: Noisy returns - MSFT 10th September, 2010



A figure of the raw noisy returns that were used to compute the set of pre-averaged returns that are higher than the threshold (red dotted line) in Figure 2.1. The largest noisy return in absolute value, from the pre-averaged returns larger than the threshold, is identified (the black circle) and removed.

3 Data description

In this section, we describe our data set and explain where it comes from.

We use the Limit Order Book System - The Efficient Reconstructor (LOBSTER) data set. LOBSTER [1] is an online limit order book tool designed to provide researchers with high-quality limit order book data for the entire universe of Nasdaq traded stocks.

The data set extracted covers the period from the 27^{th} of June 2007 to the 1^{st} of March 2015, which includes almost 2,000 trading days per security. In order to compare our results with the Christensen paper, we analyze the same 30 constituent stocks of the Dow-Jones Industrial Average index (included are both hidden and visible executed orders).

One issue that arises with comparing results using data from different providers, is that different measuring and cleaning techniques are used on the raw, order-level data, prior to any application of pre-averaging techniques (described in Section 2.4). For example, there are often revisions to old data sets that often result in some significant outliers being removed, such that jumps identified in one sample would not have been identified as a jump in another. This can result in significantly different results depending on the degree to which the cleaning techniques differ.

In addition, there is the much more significant issue of which exchange the data was recorded from. The constituents of the Dow Jones Industrial Average (DJIA) are traded in many markets and often in different denominations, world-wide. Ensuring that the data is from the same market would strengthen the results.

[1] see lobsterdata.com

4 Results

4.1 The level of jump variation

In this section, we use the pre-averaged RV^* and BV^*_τ estimators, as described in Section 2, to study the role of the jump component in the price process of a sample of US equity stocks and indices (Tables 4.1 and 4.2). In addition, we compare the tick-level derived estimates of the jump variation with results from 5- and 15-minute sampled data using realized volatility estimators RV and BV.

To begin our investigation, we look at individual stocks before discussing the equity indices. Table 4.1 displays the results for the constituents of the DJIA. The first column contains the ticker of each stock, for example the first entry is AA, the ticker for American Airlines. The second column, with heading N, shows the average daily number of executed orders. There is clearly a large variation across the stocks, with the most active instrument (APPL with 61,584 trades per day) trading at an average rate of more than ten times that of the lowest (TRV with an average of 5,872). Note that this is not the same as daily volume, as we do not use information on the size of trades in this study. The third column contains the estimated noise-to-signal ratio of the observed price process.

Focusing on the tick frequency columns, and in particular JV, the average level of the jump variation is 3.1%. This is higher than found in the Christensen, Oomen and Podolskij (2014a) paper, but not as large as found otherwise empirically (Table 1.1) in previous studies. There is a large range of JV, from -3.8% to 7.8%, derived from the instruments with the highest and lowest number of cleaned trades, respectively. The instruments with less than 10,000 daily, cleaned trades, all have more than 5% jump variation. In contrast, from the equity stocks with more than 30,000 cleaned daily trades, only one has more than 1% JV.

The level of the jump variation at the 15-minute sampling frequency averages 10.8% with a range from 9.0% to 12.9%. In comparison, the measures derived from the 5-minute sampling frequency have an average of 7.4% with a range from 6.1% to 9.8%. All instruments have a lower average jump variation at the 5- compared with the 15-minute frequency. Similarly, on an instrument level, the level of jump variation at the order-level is consistently below the 5-minute sampled estimates.

Overall, the realized volatility estimates are at a similar level for each instrument, whether estimated using tick-level or 15-minute data. There are a few notable exceptions, like XOM, where the tick level estimate is more than double that of both the 5- and 15-minute frequencies. This indicates that the real difference in JV estimates lies in the bipower variation estimators of the integrated variance.

Moving onto the equity indices, both instruments are liquid, the SPY in particular with an average of more than 170,000 trades per day. This is more than for

Tab. 4.1: Estimates of the jump variation component, using order-level returns

			Tick frequency			5-min frequency			15-min frequency		
DJIA constituents	N	$\hat{\gamma}$	RV^*	BV_τ^*	JV	RV	BV	JV	RV	BV	JV
AA	14,534	0.46	39.6	39	4.0	40.0	38.7	8.8	39.0	37.1	9.4
APPL	61,584	0.30	29.2	29.6	-3.8	28.6	28.0	6.2	28.4	26.8	10.6
AXP	15,847	0.24	36.4	35.6	3.6	36.5	35.2	6.6	36.0	33.9	9.8
BA	9,252	0.20	25.4	24.9	5.2	25.7	24.8	7.1	25.7	24.0	11.0
BAC	43,574	0.60	47.1	47.0	3.0	47.5	45.3	8.4	47.0	44.2	11.0
CAT	13,404	0.18	30.6	30.1	2.9	30.6	29.8	6.1	30.4	29.0	9.6
CSCO	35,458	0.60	30.8	26.3	0.9	26.9	25.9	8.6	26.4	25.0	10.0
CVX	17,667	0.22	24.7	24.6	1.8	24.7	24.1	6.6	24.1	23.0	9.0
DD	10,121	0.24	26.5	26.2	3.3	26.9	26.0	6.4	26.7	25.2	9.1
DIS	13,244	0.30	24.8	24.3	4.0	25.1	24.0	7.5	24.9	23.4	9.7
GE	28,344	0.53	30.5	30.2	2.7	30.5	29.2	8.7	30.5	28.3	10.9
HD	17,602	0.31	27.8	27.2	3.9	27.7	26.8	7.1	27.3	25.5	10.5
HPQ	19,707	0.33	28.1	27.6	3.0	28.5	27.3	7.6	28.0	26.0	11.7
IBM	10,943	0.22	21.4	21.1	3.7	21.3	20.5	7.7	20.8	19.5	10.9
INTC	38,951	0.61	27.2	27.1	0.0	29.5	26.5	8.0	29.0	24.9	11.0
JNJ	15,615	0.33	16.3	16.0	3.7	16.4	15.6	9.1	16.1	15.0	12.5
JPM	44,646	0.34	38.9	38.8	0.5	38.8	37.6	6.1	38.3	36.1	10.3
KO	13,566	0.34	18.0	17.7	2.8	18.4	17.2	9.2	17.9	16.4	12.9
MCD	11,171	0.23	19.2	18.8	4.0	19.2	18.4	6.9	18.6	17.3	11.3
MMM	6,555	0.23	21.7	21.0	5.1	21.5	20.8	7.5	20.9	19.6	10.1
MRK	15,993	0.35	24.4	23.8	4.4	25.4	24.0	8.7	24.3	22.4	11.6
MSFT	45,805	0.57	24.2	24.2	0.2	24.6	23.4	8.1	24.0	22.2	11.4
PFE	20,732	0.55	22.7	22.3	3.2	22.8	21.7	9.8	21.9	20.7	11.0
PG	15,198	0.30	19.4	18.1	3.2	17.4	16.6	8.4	16.9	15.6	11.7
T	19,836	0.45	23.2	22.9	3.1	23.6	22.8	9.1	22.4	21.1	11.1
TRV	5,872	0.25	29.2	27.2	7.4	29.8	28.3	8.7	29.0	26.0	12.3
UTX	8,843	0.23	22.7	22.0	6.1	23.1	22.1	7.6	22.5	21.1	10.4
VZ	15,739	0.38	22.5	22.1	3.4	22.8	21.8	8.6	22.2	20.8	11.2
WMT	19,827	0.34	19.3	19.1	3.0	19.4	18.6	8.1	18.8	17.5	12.1
XOM	33,445	0.15	46.3	46.3	0.2	22.9	22.3	6.1	22.3	21.2	9.8
Average	21,436	0.35	27.3	26.7	3.0	26.5	25.4	7.8	26.0	24.3	10.8
Christensen et al. Average	35,541	0.41	31.6	31.4	1.3	31.7	30.6	7.3	30.8	28.9	11.9

Estimated JV from the corresponding BV_τ^* and RV^* estimates at tick-frequency across the 30 DJIA index constituents on dates 27^{th} of June 2007 through to the 1^{st} of March 2015. Also shown are JV estimates, from the corresponding RV and BV estimates, at the 5- and 15-minute sampling frequencies. N is the average number of daily (cleaned) trades, and $\hat{\gamma}$ is the (estimated) noise-ratio parameter, given by $\hat{\gamma} = \sqrt{\frac{N \cdot \omega \hat{A}_C{}^2}{\int_0^1 \sigma_s^2 ds}}$. RV^* is the noise- and outlier-robust realized volatility estimator, given by equation 11 and BV_τ^* is the bipower variation derived from filtered returns used on equation 12. JV is the jump variation contribution from the total variation, equation 7. The RV and BV, at the 5- and 15-minute sampling frequencies use the standard methods described in equations 1 and 8, respectively.

any individual constituent of the DJIA. Both equity indices have negative jump variation derived from tick-level data, similar to Christensen, Oomen and Podolskij (2014a). Similarly, the level of jump variation increases with a decreasing sampling frequency.
A general pattern emerges - the higher the number of trades, measured by the average number of daily observations, the lower the relative contribution of the jump variation. The most liquid instruments, such as Apple Inc., or the ETF of the S&P 500 have a negative level of jump variation. This is counter-intuitive, to some extent and may suggest that the method of removing outliers is too rigorous. This could be because the calibration of the threshold τ is based on simulated data assuming 40,000 trades a day. This is higher than the average in our data set, which is just 21,000. Instruments with very high levels of daily trades could possibly have an overestimated bipower variation. The Christensen, Oomen and Podolskij (2014a) paper also found negative JV for the majority of the most liquid instruments in their sample. This could be due to the overestimation of the bipower variation measure (BV^*_τ) for instruments with a larger than average number of daily trades; the threshold τ is dependent on the number of daily trades N in addition to the pre-averaging horizon θ (which was calibrated on a simulation with 40,000 trades).

The general result may be explained by a more fundamental difference. While the data set used in the Christensen, Oomen and Podolskij (2014a) paper covered approximately the same period as our study, the daily frequency of observations reported in the aforementioned paper was, on average, around 50% higher. Hence, the difference could be simply down to more stringent data-cleaning algorithms used on the limit order book data: the lower frequency of our data set could account for the higher level of JV. The results from our analysis certainly show this. Despite these differences, a clear trend can be seen: the higher the average number of daily trades, the lower the level of JV. Even within each study, the instruments with the highest number of average daily trades have the lowest levels of jump variation, some even negative.

Looking back to the empirical findings shown in Table 1.1, the level of jump variation found in this study, is still lower than what has been identified previously. The lower estimates of Barndorff-Nielsen and Shephard (2004a) were found to be slightly higher, but those results were from FX-rates that cannot be reliably compared with equities and index trackers.

Tab. 4.2: Estimates of the jump variation component, using order-level returns

			Tick frequency			5-minute frequency			15-min frequency		
Equity indices	N	$\hat{\gamma}$	RV^*	BV^*_τ	JV	RV	BV	JV	RV	BV	JV
QQQ	22,442	0.30	12.9	13.0	-0.4	13.1	12.8	5.8	13.1	12.7	7.5
SPY	173,166	0.50	17.2	17.4	-1.3	17.3	16.9	6.5	17.1	16.4	7.7
Average	97,804	0.40	15.1	15.2	-0.9	15.2	14.9	6.1	15.1	14.5	7.6
Christensen et al. Average	81,052	0.36	22.0	22.1	-1.3	21.9	21.5	3.9	21.7	20.7	8.5

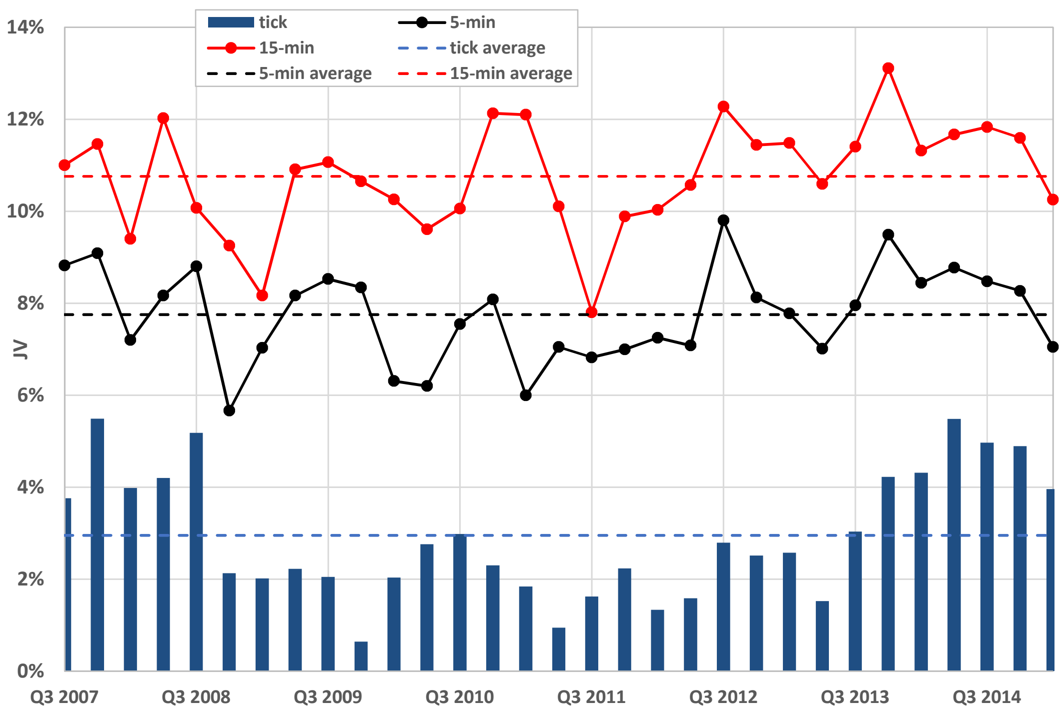
JV and corresponding BV^*_τ and RV^* estimates at tick-frequency across 2 index trackers on dates 27^{th} of June 2007 through to the 1^{st} of March 2015. Also shown are JV estimates, from the corresponding RV and BV estimates, at the 5- and 15-minute sampling frequencies. N is the average number of daily (cleaned) trades, and $\hat{\gamma}$ is the (estimated) noise-ratio parameter, given by $\hat{\gamma} = \sqrt{\frac{N \cdot \omega_{\hat{AC}}{}^2}{\int_0^1 \sigma_s^2 ds}}$. RV^* is the noise- and outlier-robust realized volatility estimator, given by equation 11 and BV^*_τ is the bipower variation derived from filtered returns used on equation 12. JV is the jump variation contribution from the total variation, equation 7. The RV and BV at the 5- and 15-minute sampling frequencies use the standard methods described in equations 1 and 8 respectively.

4.2 The level of jump variation across time

As part of our analysis, we also look at the quarterly jump variation estimates for the three sampling frequencies over the whole 8-year period (Figures 4.1 and 4.2). The results are stable for the equities covered by the DJIA (Figure 4.1). The level of JV varies from under 1% in the last quarter of 2009 to almost 6% in the 4^{th} quarter of 2007. There are some notable exceptions, for example the 3^{rd} quarter of 2008 representing the credit-crunchcrisis, where the average level of jump variation is almost 5% of the total variation. For each quarter, the level of jump variation at the tick-level is consistently and markedly lower than for the lower frequencies.

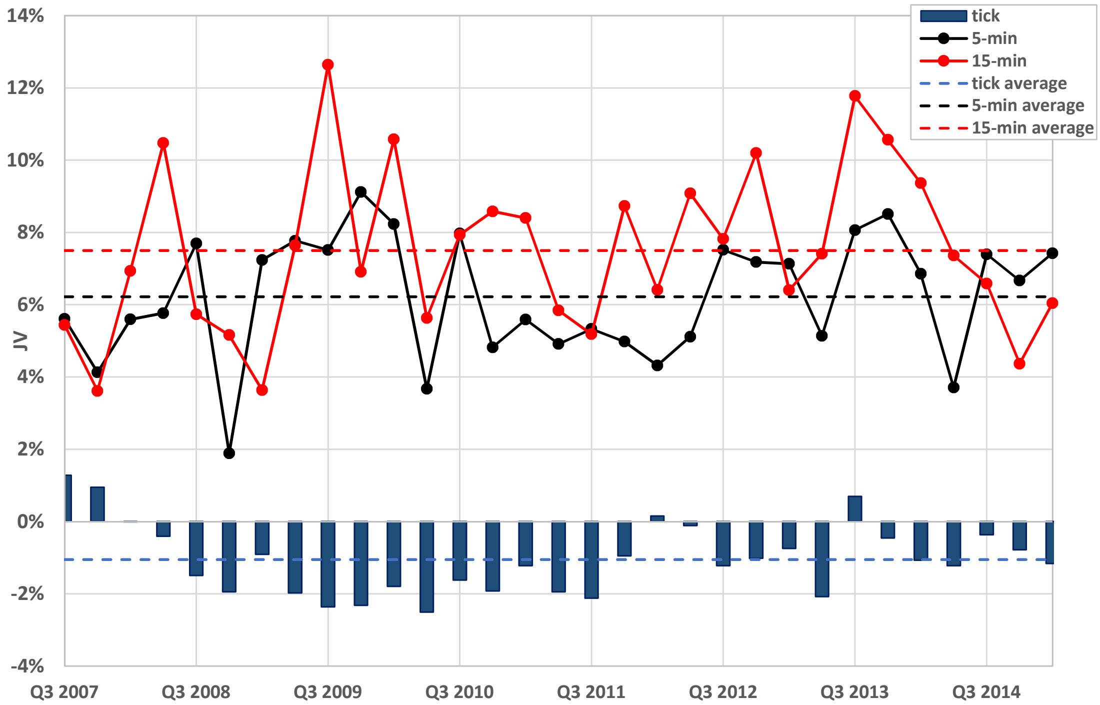
The level of jump variation of the index trackers across time (shown in Figure 4.2) is slightly more stable than for DJIA constituents when comparing with tick-level data. This may be because there were only two equity indices in the study, but 30 stocks. However, this is not matched by the 5- and 15-minute sampled results, where the JV estimates of the indices are more volatile across time. Contrastingly, there are a few quarters (mostly in the years of the financial crisis - 2008 and 2009) where the 5-minute JV of the indices exceeds that of the 15-minute estimate. That said, all 5- and 15-minute estimates of JV are positive for all instruments.

Fig. 4.1: Equity stocks: jump proportion, quarter-by-quarter



Quarterly JV over 8-year period with different sampling frequencies - DJIA constituents.

Fig. 4.2: Equity indices: jump proportion, quarter-by-quarter



Quarterly JV over 8-year period with different sampling frequencies - index tracker funds.

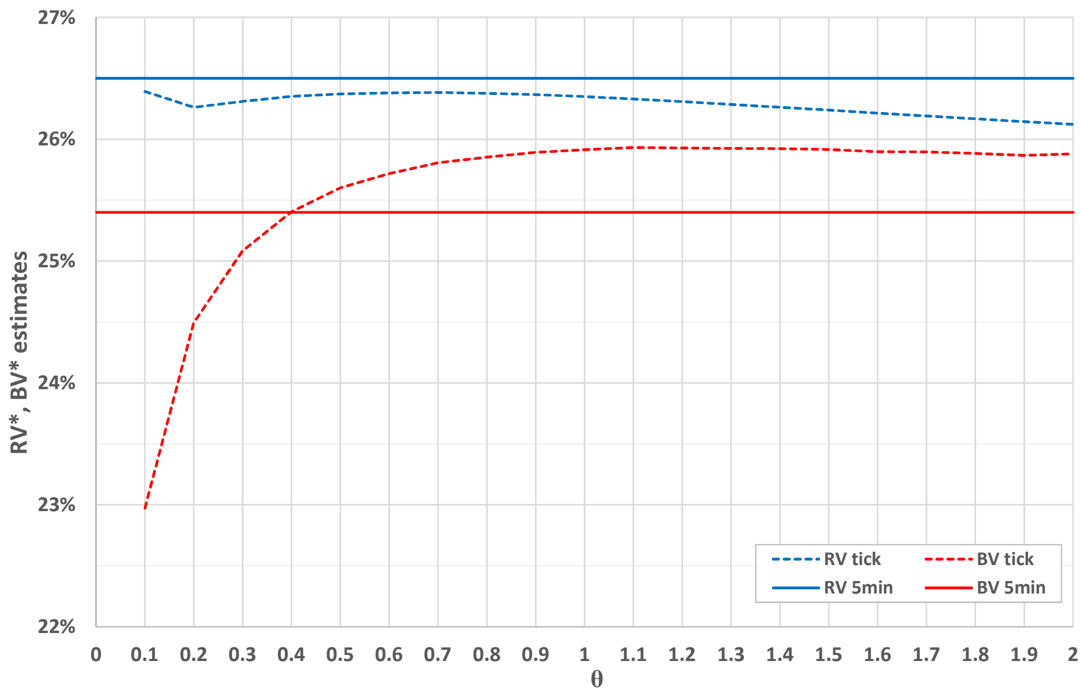
4.3 The suitability of the pre-averaging horizon θ

In order to gauge the suitability of the pre-averaging horizon θ used in the study, a look at how the variation estimates vary across pre-averaging horizons is discussed in this subsection.

Figure 4.3 shows how estimates of RV^* and BV_τ^*, averaged over all constituents of the DJIA, vary for pre-averaging horizons θ in the range of $[0.1, 2]$. The first thing to note is that the realized volatility estimate at the 5-min sampling frequency is consistently higher than that found from tick-level data. The bipower variation is lower at the 5-minute sampled RV measure, for $\theta > 0.4$. In the range of $\theta \in [0.5, 2]$, the RV^* estimate is decreasing with an increasing pre-averaging horizon. On the other hand, BV_τ^* measure is increasing, then it is mostly constant for $\theta > 1$.

From the figure, it is clear that the average level of RV^* and BV_τ^* does not vary much for levels of θ greater than 0.6.

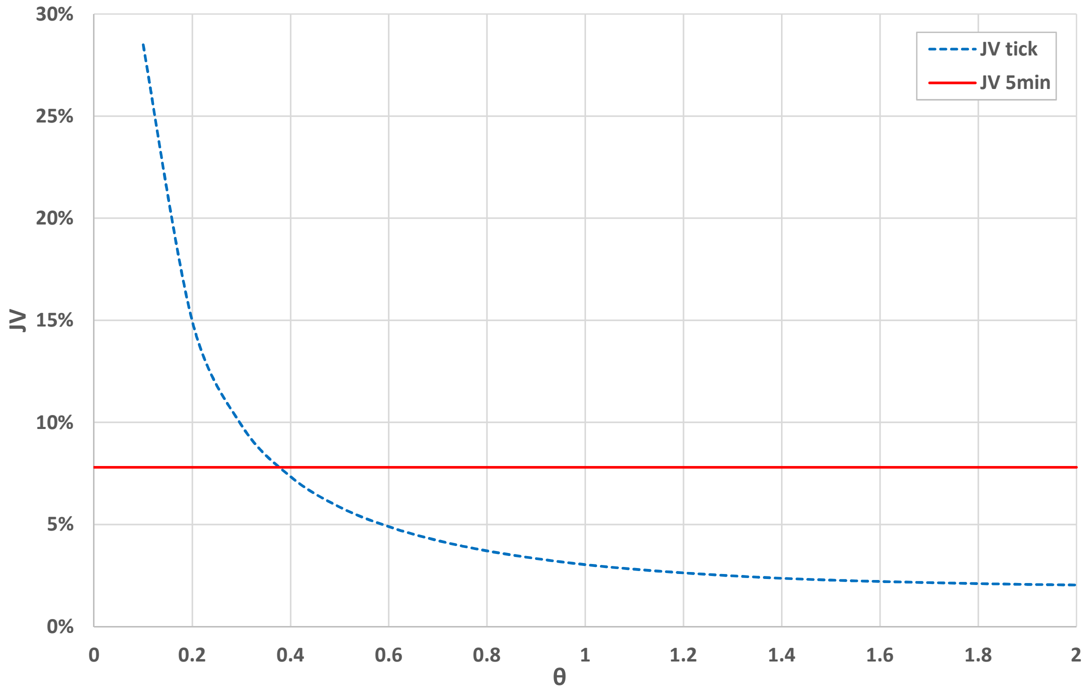
Fig. 4.3: RV^* and BV_τ^* θ-signature plot across all DJIA constituents



RV^* and BV_τ^* θ-signature plot for the DJIA-constituents. The RV^* estimates are show with the dotted red line, while the BV_τ^* estimates are in blue. The 5-minute sampled estimates are shown by the solid lines.

Figure 4.4 displays how the average JV estimate varies for different pre-averaging horizons, compared with the 5-minute sampled estimate. For small pre-averaging horizons, the level of JV is large as the effects of noise begin to upwardly bias the JV estimate. Larger pre-averaging horizons decrease the estimate of JV, but the effect diminishes. The average estimate of JV for the DJIA constituents is 2%, when $\theta = 2$. This is closer to the 1% level found in Christensen, Oomen and Podolskij (2014a), but still almost double the contribution to total volatility.

Fig. 4.4: JV^* θ-signature plot across all DJIA-constituents



JV^* θ-signature plot for the DJIA-constituents. The blue dotted line shows how the level of JV varies over the pre-averaging horizon θ for tick-level data. The red solid line shows JV estimated from 5-minute sampled data, as a comparison.

5 Conclusion

This paper attempts to replicate results found in Christensen, Oomen and Podolskij (2014a). Christensen, Oomen and Podolskij (2014a) found that the level of jump variation estimated using high-frequency data is 1.3%. This contrasts with the 10% level found by previous studies (see Table 1.1). Tick-level, high frequency data has been until recently, mostly unused in jump variation estimations due to the inability of measures to handle the noisy, outlier-ridden data. Utilizing novel econometric techniques developed by Christensen, Oomen and Podolskij (2014a), on the same set of financial instruments, across a slightly expanded time period, we are able to clean the noise effects away and find that the level of jump variation is 3.0%. This is higher than that found in the Christensen, Oomen and Podolskij (2014a) paper, but still considerably lower than previous empirical studies. The difference between the estimates of JV using tick-level data and 5-minute and higher sampled data is explained, in part by the sampling frequency. 5-minute sampled data picks up smooth, but fast price movements as jumps, which JV estimates using tick level data characterizes as bursts of volatility.

We conclude that there is a sizable difference in JV estimates across instruments from the same asset class, with the average number of daily trades seemingly a principal contributor to the differences. Instruments averaging more than 30,000 daily trades averaged a jump variation of 0.2%, in line with results from Christensen, Oomen and Podolskij (2014a) for instruments with similar trading volumes. This aspect may resolve the difference in results with the Christensen paper (where the average number of daily returns was 35,000).

Overall, we find that our results are consistent with Christensen, Oomen and Podolskij (2014a). The role of the jump variation is indeed much lower than previously thought. This implies that jump-diffusion models may not fit with empirical data.

Literatur

Aït-Sahalia Y., 2004, "Disentangling diffusion from jumps," *Journal of Financial Economics*, 74(3), 487–528.

Aït-Sahalia Y., and Jacod J., 2009a, "Estimating the degree of activity of jumps in high frequency data," *Annals of Statistics*, 37(5), 2202–2244.

Aït-Sahalia Y., and Jacod J., 2009b, "Testing for jumps in a discretely observed process," *Annals of Statistics*, 37(1), 184–222.

Alizadeh S., Brandt M. W. , and Diebold F. X., 2002, "Range-based estimation of stochastic volatility models," *Journal of Finance*, 57(3), 1047–1092.

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," *Review of Economics and Statistics*, 89(4), 701–720.

Andersen T. G., Bollerslev T., and Diebold F. X., 2010, "Parametric and nonparametric volatility measurement," *Handbook of Financial Econometrics*, ed. by L. P. Hansen, and Y. Aït-Sahalia. North-Holland, pp. 67–138.

Andersen T. G., Bollerslev T., and Huang X., 2011, "A reduced form framework for modeling volatility of speculative prices based on realized variation measures," CREATES Research Papers 2007–14, School of Economics and Management, University of Aarhus.

Andersen T. G., Dobrev D., and Schaumburg E., 2009, "Jump-robust volatility estimation using nearest neighbor truncation," *Journal of Econometrics*, 169(1), 75–93.6

Andersen T. G., Dobrev D., and Schaumburg E., 2013, "A robust neighborhood truncation approach to estimation of integrated quarticity," *Econometric Theory*, 30(01), 3–59.

Andersen T.G., Benzoni L., and Lund J., 2002, "An empirical investigation of continuous-time equity return models," *Journal of Finance*, 57(4), 1239–128.

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.

Bachelier L., 1900, "Théorie de la spéculation.," Gauthier-Villars.

Back K., 1991, ”Asset prices for general processes,” *Journal of Mathematical Economics*, 20(4), 371–395.

Bakshi G. S., Cao C. , and Chen Z., 1997, ”Empirical performance of alternative option pricing models,” *Journal of Finance*, 52(2), 2003–2049.

Ball C. A., and Torous W. N., 1983, ”A simplified jump process for common stock returns,” *Journal of Financial and Quantitative Analysis*, 18(1), 53–65.

Barndorff-Nielsen O. E., and Shephard N., 2002a, ”Econometric analysis of realised volatility and its use in estimating stochastic volatility models,” *Journal of the Royal Statistical Society*, B 64, 253–280.

Barndorff-Nielsen O. E., and Shephard N., 2002b, ”Estimating quadratic variation using realised volatility,” *Journal of Applied Econometrics*, 17, 457–477.

Barndorff-Nielsen O. E., and Shephard N., 2004a, ”Power and bipower variation with stochastic volatility and jumps,” *Journal of Financial Econometrics*, 2(1), 1–48.

Barndorff-Nielsen O. E., and Shephard N., 2004b, ”Econometrics of testing for jumps in financial economics using bipower variation,” *Journal of Financial Econometrics*, 4(1), 1–30.

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., Graversen S. E., Jacod J., Podolskij M., and Shephard N., 2006, ”A central limit theorem for realized power and bipower variations of continuous semimartingales,” in From Stochastic Calculus to Mathematical Finance: The Shiryaev Festschrift ed. by Y. Kabanov, R. Lipster.

Bates D. S., 2000, ”Post-‘87 crash fears in the S&P500 futures option market,” *Journal of Econometrics*, 94, 181–238

Black F., 1976, ”Studies of stock market volatility changes,” *in Proceedings of the American Statistical Association*, pp. 177–181. Business and Economic Statistics Section.

Bollerslev T., and Zhou H., 2002, ”Estimating stochastic volatility diffusion using conditional moments of integrated volatility,” *Journal of Econometrics*, 109(1), 33–65.

Bollerslev T., Kretschmer U., Pigorsch C., and Tauchen G., 2009, ”A discrete-time model for daily S&P500 returns and realized variations: jumps and leverage effects,” *Journal of Econometrics*, 150(2), 151–166.

Bollerslev T., Law T. H., and Tauchen G., 2008, ”Risk, jumps, and diversification,” *Journal of Econometrics*, 144(1), 234–256.

Chernov M., Gallant A. R. , Ghysels E., and Tauchen G., 2003, ”Alternative models for stock price dynamics,” *Journal of Econometrics*, 116(1-2), 225–257.

Christensen K., Oomen R. C. A., and Podolskij M., 2014a, ”Fact or fiction: Jumps at ultra high frequency,” *Journal of Financial Economics* 114(3) 576–599, School of Economics and Management, University of Aarhus.

Christensen K., Oomen R. C. A., and Podolskij M., 2014b, ”Fact or fiction: Appendix,” Working paper. Available at http://ssrn.com/abstract=1848774.

Corsi F., and Renò R., 2012, ”Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling,” *Journal of Business and Economic Statistics*, 30(3), 368–380

Corsi F., Pirino D., and Renò R., 2010, ”Threshold bipower variation and the impact of jumps on volatility forecasting,” *Journal of Econometrics*, 159(2), 276–288.

Cox, J. C., and S. A. Ross, 1976, ”The valuation of options for alternative stochastic processes,,” *Journal of Financial Economics*, 3(1-2), 145–166.

Delbaen F., and Schachermayer W., 1994, ”A general version of the fundamental theorem of asset pricing,” *Mathematische Annalen*, 300, 463–520.

Diebold F. X., and Strasser G. H., 2008, ”On the correlation structure of microstructure noise in theory and practice,” PIER Working Paper Archive 08–038, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Easley D., de Prado M. M. L., and O’Hara M., 2011, ”The microstructure of the flash crash: Flow toxicity, liquidity crashes and the probability of informed trading,” *Journal of Portfolio Management*, 37(2), 118–128.

Eraker B., 2004, ”Do stock prices and volatility jump? Reconciling evidence from spot and option prices,” *Journal of Finance*, 59(3), 1367–1403.

Eraker B., Johannes M., and Polson N., 2003, ”The impact of jumps in volatility and returns,” *Journal of Finance*, 58(3), 1269–1300.

Hansen P. R., and Lunde A., 2006, "Realized variance and market microstructure noise," *The Journal of Business and Economic Statistics*, 24(2), 127–161.

Hautsch N., and Podolskij M., 2010, "Pre-averaging based estimation of quadratic variation in the presence of noise and jumps; theory, implementation and empirical evidence," SFB 649 Discussion Papers SFB649DP2010-038, Humboldt University, Germany.

Huang X., and Tauchen G., 2005, "The relative contribution of jumps to total price variance," *Journal of Financial Econometrics*, 3(4), 456–499.

Jacod J., Li Y., Mykland P. A., Podolskij M., and Vetter M., 2009, "Microstructure noise in the continuous case - the pre-averaging approach," *Stochastic Processes and their Applications*, 119(7), 2249–2276.

Jarrow R. A., and Rosenfeld E. R., 1984, "Jump risks and the intertemporal capital asset pricing model," *Journal of Business*, 57(3), 337–351.

Jiang G. J., and Oomen R. C. A., 2007, "Estimating latent variables and jump diffusion models using high-frequency data," *Journal of Financial Econometrics*, 5(1), 1–30.

Maheu J. M., and McCurdy T. H. 2004, "Dynamic asset allocation with event risk," *Journal of Finance*, 59(2), 755–793

Mancini C., 2009, "Estimation of the characteristics of jump of a general Poisson-diffusionmodel," *Scandinavian Actuarial Journal*, 2004(1), 42–52.

Merton R. C., 1976, "Option pricing when underlying stock returns are discontinuous," *Journal of Financial Economics*, 3(1–2), 125–144.

Oomen R. C. A., 2006, "Comment on 2005 JBES invited address "Realized variance and market microstructure noise" by Peter R. Hansen and Asger Lunde," *Journal of Business and Economic Statistics*, 24(2), 195–202.

Patton A. J., and Sheppard N., 2013, "Good volatility, bad volatility signed jumps and the persistence of volatility," Working paper. Available at ssrn.com/abstract=1943825.

Podolskij M., and Vetter M., 2009a, "Bipower-type estimation in a noisy diffusion setting," *Stochastic Processes and their Applications*, 119(9), 2803–2831.

Podolskij M., and Vetter M., 2009b, "Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps," *Bernoulli*, 15(3), 634–658.

Press S. J., 1967, "A compound events model for security prices," *Journal of Business*, 40(3), 317–335

Tauchen G. E., and Zhou H., 2011, "Realized jumps on financial markets and predicting credit spreads," *Journal of Econometrics*, 160(1), 102–118.

Todorov V., 2009, "Estimation of continuous-time stochastic volatility models with jumps using high-frequency data," *Journal of Econometrics*, 148(2), 131–148.