Fact or friction: Jumps at ultra high frequency

Kim Christensen*  Roel Oomen†  Mark Podolskij‡,*

*CREATES  †Deutsche Bank  ‡University of Heidelberg

January 24, 2011
Outline I

1 Introduction

2 Theoretical setup
   - Semimartingale framework
   - Estimating quadratic variation
   - Microstructure noise and outliers
   - Estimating variance/covariance matrix

3 Simulation study
   - Simulation details
   - Results: Part I
   - Threshold estimation
   - Results: Part II
In recent years, high-frequency jump estimation has attracted increasing interest (e.g., Andersen, Bollerslev and Diebold 2007; Barndorff-Nielsen and Shephard 2004, 2006; Corsi and Renó 2009; Huang and Tauchen 2005).

Jumps appear to be frequent and account for a significant proportion of total return variation (ranging about 5% – 15%).

Most studies use sparsely sampled data, for example 5-minute data.

We investigate the importance of the jump component with noise- and outlier-robust estimators using ultra high-frequency data.

We find much less evidence of jumps, i.e. a substantially smaller jump proportion and fewer significant jump days.
Semimartingale framework

- We assume that a log-price $X_t$ at time $t$ is, potentially, of the form

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

where

- $X_0$ is the initial price, $\mu$ is a drift term, $\sigma$ is a (stochastic) volatility process, $W$ is a Brownian motion, while $N_t^J$ and $(J_i)_{i \geq 0}$ represent the total number and sizes of jumps up to time $t$.

- $X$ represents an underlying “efficient” price that would prevail in the absence of microstructure noise and outliers.
We normalize time to the unit interval, \( t \in [0, 1] \).

The quadratic variation of \( X \)

\[
[X]_1 = \int_0^1 \sigma_s^2 \, ds + \sum_{i=1}^{N_i^J} J_i^2 \equiv \text{IV} + \text{JV}.
\]

The question we investigate is how important the jump variation (JV) is relative to the integrated variance (IV).

Let the observation times of \( X \) be equidistant time points \( t_i = i/N \), for \( i = 0, 1, \ldots, N \) Then, we compute log-returns by

\[
\Delta_i^N X = \frac{X_{i/n}}{n} - \frac{X_{i-1/n}}{n}.
\]
**Realised variance and bipower variation**

- The realised variance and bipower variation are defined by

\[
RV_N[X] = \sum_{i=1}^{N} |\Delta^N_i X|^2 \quad BV_N[X] = \frac{N}{N-1} \frac{\pi}{2} \sum_{i=2}^{N} |\Delta^N_i X||\Delta^N_{i-1} X|.
\]

- Moreover, it holds that

\[
RV_N[X] \xrightarrow{p} [X]_1 \quad BV_N[X] \xrightarrow{p} IV.
\]

- That is, bipower variation is a jump-robust estimator of the integrated variance, and \( RV_N[X] - BV_N[X] \xrightarrow{p} JV \).
A high-frequency jump test

- Assume there are no jumps in $X$, i.e. $N^J_t \equiv 0$. Then, the following CLT holds

$$
\frac{N^{1/2}(RV_N[X] - BV_N[X])}{\sqrt{0.609 \int_0^1 \sigma_s^4 ds}} \xrightarrow{d} N(0, 1).
$$

- Basis for performing non-parametric tests of the existence of jumps in the absence of microstructure noise and outliers.

- A feasible version is achieved by plugging in a consistent estimator of the integrated quarticity, $\int_0^1 \sigma_s^4 ds$.

- The t-statistic can be further transformed using the delta method to improve finite sample properties.
Microstructure noise and outliers

- In practice, microstructure noise (e.g., bid-ask spreads and price discreteness) cloaks the true log-price $X$.
- Moreover, data are contaminated with outliers (e.g., due to misplaced decimal points, errors in the data feeds etc.).
- Outliers can be hard to filter out systematically.
- We model the observed log-price $Y$ as

$$Y_{iN} = X_{iN} + u_{iN} + 1_{i \in A_{N}^O}O_i$$

- $u$ is an i.i.d. microstructure noise process with $E(u) = 0$ and $\text{var}(u) = \omega^2$. Moreover, $u$ is independent of $X$, i.e. $u \perp \perp X$. 
\((O_i)_{i \geq 0}\) are non-zero random variables, which generates the sizes of the outliers.

\(A^O_N\) is a random set, which holds the appearance times of outliers. We assume \(A^O_N\) is a.s. finite and model it by

\[
A^O_N = \left\{ \frac{[N \times T^O_i]}{N} : 0 \leq T^O_i \leq 1, i \geq 1 \right\}
\]

\((T^O_i)_{i \geq 0}\) are the jump times of another counting process \(N^O_t\), where \(N^O_t\) is independent of \(N^J_t\).

The independence between \(N^J_t\) and \(N^O_t\) implies that the processes have no common jumps.

Thus, observing both a jump and an outlier in \(Y\) over a small time interval is very unlikely.
Case I: The noiseless case

Assume first that $u \equiv 0$, i.e. there is no noise, but there could be outliers in the data, $Y_{i/N} = X_{i/N} + 1_{i \in A_N^O} O_i$, $i = 0, 1, \ldots, N$.

Theorem I

In the absence of microstructure noise but presence of outliers, the following convergence in probability holds

$$RV_N[Y] \xrightarrow{p} [X]_1 + 2 \sum_{i=1}^{N_1^O} O_i^2$$

$$BV_N[Y] \xrightarrow{p} IV + \frac{\pi}{2} \sum_{i=1}^{N_1^O} O_i^2$$
Thus, neither estimator is consistent for the object, they are designed to estimate. Moreover, even in absence of jumps

$$RV_N[Y] - BV_N[Y] \xrightarrow{p} (2 - \pi/2) \sum_{i=1}^{N_1^O} O_i^2 > 0.$$  

Thus, a jump test based $RV_N[Y] - BV_N[Y]$ will reject the null with probability converging to 1, also under the null of no jumps!

To estimate $[X]_1$ and the IV, we use a third estimator, the (subsampled) QRV of Christensen, Oomen and Podolskij (2010):

$$QRV_N[Y] \equiv \alpha' QRV_N(m, \bar{\lambda})[Y],$$
Here, \( \overline{\lambda} = (\lambda_1, \ldots, \lambda_k) \) with \( \lambda_j \in [0, 1) \) is a vector of quantiles, \( \alpha = (\alpha_1, \ldots, \alpha_k) \) are quantile weights with \( \alpha_j \geq 0, \sum \alpha_j = 1 \) and \( \text{QRV}_N(m, \overline{\lambda})[Y] \) is a \((k \times 1)\) vector with \( j \)th entry equal to

\[
\text{QRV}_N(m, \lambda_j)[Y] = \frac{1}{N - m} \sum_{i=1}^{N-m} \frac{q_i(m, \lambda_j)}{\nu_1(m, \lambda_j)}, \quad \text{and}
\]

\[
q_i(m, \lambda) = g^2_{\lambda m} \left( \sqrt{N} |D_{i,m} Y| \right),
\]

and where \( D_{i,m} Y = (\Delta^N_k Y)_{(i-1)m+1 \leq k \leq im} \) for \( i = 1, \ldots, n \).

Under the assumptions of Theorem I, it holds that

\[
\text{QRV}_N[Y] \overset{p}{\rightarrow} IV. \tag{1}
\]
We can further identify the jump and outlier variation by taking appropriate linear combinations of $RV_N[Y]$, $BV_N[Y]$ and $QRV_N[Y]$

$$RV_N[Y] - \left(1 - \frac{4}{\pi}\right) QRV_N[Y] - \frac{4}{\pi} BV_N[Y] \xrightarrow{p} JV$$

$$\frac{2}{\pi} (BV_N[Y] - QRV_N[Y]) \xrightarrow{p} \sum_{i=1}^{N_1^O} O_i^2$$

An application of the delta method to the joint CLT (under no noise) of $(RV_N[Y], BV_N[Y], QRV_N[Y])$ can be used to test for jumps in the presence of outliers (see paper for details).
Case II: The noise case

- We now add the noise back and consider the simultaneous impact of noise and outliers, i.e.

\[ Y_{i/N} = X_{i/N} + u_{i/N} + 1_{i \in A_{N}O_i}, \quad i = 0, 1, \ldots, N. \]

- Well-known that standard estimators based on \( Y \), e.g., \( RV_N[Y] \), \( BV_N[Y] \) or \( QRV_N[Y] \) are inconsistent under noise.

- To infer the characteristics of the underlying semimartingale, we apply the pre-averaging approach, see, e.g., Jacod, Li, Mykland, Podolskij and Vetter (2009) or Podolskij and Vetter (2009a,b).
Two ingredients are needed. First, we choose a sequence of integers

\[ K = K(N) = \theta \sqrt{N} + o(N^{-1/2}), \quad \theta > 0. \]

In the paper, we use \( K = \lceil \theta \sqrt{N} \rceil \).

The second ingredient is a pre-averaging function \( g \), which has to satisfy some technical conditions (see paper).

Throughout, we work with \( g(x) = \min(x, 1 - x) \).

Associated with \( g \) are some normalizing constants:

\[
\psi^K_1 = K \sum_{j=1}^{K} \left( h \left( \frac{j}{K} \right) - h \left( \frac{j-1}{K} \right) \right)^2, \quad \psi^K_2 = \frac{1}{K} \sum_{j=1}^{K-1} h^2 \left( \frac{j}{K} \right)
\]
We then pre-average noisy returns

$$\bar{Y}_i^N = \sum_{j=1}^{K} g \left( \frac{j}{K} \right) \Delta_{i+j}^N Y.$$

An equivalent representation

$$\bar{Y}_i^N = \frac{1}{K} \sum_{j=K/2}^{K-1} Y_{i+j}^N - \frac{1}{K} \sum_{j=0}^{K/2-1} Y_{i+j}^N,$$

with $K$ even and $g(x) = \min(x, 1-x)$.

Hence, the term “pre-averaging”.
Pre-averaged RV, BV and QRV

- We define noise-robust estimators:

\[
RV_{\bar{N}}^*[Y] = \left[ \frac{N}{N - K + 2} \frac{1}{K \psi_2^K} \sum_{i=0}^{N-K+1} |\bar{Y}_i^N|^2 \right] - \frac{\psi_1^K}{\theta^2 \psi_2^K} \hat{\omega}^2,
\]

\[
BV_{\bar{N}}^*[Y] = \left[ \frac{N}{N - 2K + 2} \frac{1}{K \psi_2^K \mu_1} \sum_{i=0}^{N-2K+1} |\bar{Y}_i^N||\bar{Y}_{i+K}^N| \right] - \frac{\psi_1^K}{\theta^2 \psi_2^K} \hat{\omega}^2,
\]

where \(\hat{\omega}^2\) is a consistent estimator of \(\omega^2\).

- In the paper, \(\omega^2\) is estimated following Oomen (2006)

\[
\hat{\omega}^2_{AC} = -\frac{1}{N-1} \sum_{i=2}^{N} \Delta_i^N Y \Delta_{i-1}^N Y \overset{p}{\to} \omega^2.
\]
Construction of $QRV^{*}_{N}[Y]$ is slightly more involved.

$$QRV^{*}_{N}[Y] \equiv \alpha^{'} QRV^{*}_{N}(m, \bar{\lambda})[Y],$$

where $\bar{\lambda}$ and $\alpha$ are as above, and the $j$th element of $QRV^{*}_{N}(m, \bar{\lambda})[Y]$ is given by:

$$QRV^{*}_{N}(m, \lambda_{j})[Y] = \frac{1}{\theta \psi_{2}(N - m(K - 1) + 1)} \sum_{i=0}^{N-m(K-1)} \frac{q^{*}_{i}(m, \lambda_{j})}{\nu_{1}(m, \lambda_{j})}.$$  \hspace{1cm} (2)

where

$$q^{*}_{i}(m, \lambda) = g^{2}_{\lambda m} \left( N^{1/4} |\bar{D}^{N}_{i} Y | \right),$$

and

$$\bar{D}^{N}_{i} Y = \left\{ \bar{Y}^{N}_{i+(j-1)(K-1)} \right\}_{j=1}^{m}, \quad \text{for} \quad i = 0, 1, \ldots, N - m(K - 1).$$
Theorem II

Assume that the observed log-price $Y$ obeys

$$Y_{i/N} = X_{i/N} + u_{i/N} + 1_{i \in A^O_N} O_i, \quad i = 0, 1, \ldots, N.$$ 

and $E(u^4) < \infty$. Then, it holds that

$$RV^*_N[Y] \xrightarrow{p} [X]_1$$

$$BV^*_N[Y] \xrightarrow{p} IV$$

$$QRV^*_N[Y] \xrightarrow{p} IV$$
Remarks

- In contrast to the previous results, all noise-corrected estimators are also robust to outliers!
- Intuition: With a probability "close to" one, there is at most a single outlier in the window \([i/N, (i + K)/N]\).
- The outlier influences exactly two consecutive returns with opposite sign and therefore appears with a factor \(O(|g(j/K) - g((j - 1)/K)|)\) in the construction of \(\bar{Y}_i^N\).
- But \(|g(j/K) - g((j - 1)/K)| = O(1/K)\), so outliers therefore have no impact on \(\bar{Y}_i^N\) asymptotically.
Theorem III

Assume that $N_t^j \equiv 0$, i.e. the observed log-price $Y$ is a continuous semimartingale with noise and outliers. Furthermore, we assume that $E(u^8) < \infty$. As $N \to \infty$, it holds that

$$N^{1/4} \begin{pmatrix} RV_N^*[Y] - IV \\ BV_N^*[Y] - IV \\ QRV_N^*[Y] - IV \end{pmatrix} \xrightarrow{d_s} MN(0, \Sigma^*).$$

where $\xrightarrow{d_s}$ denotes stable convergence in law and $\Sigma^*$ is the (unknown) conditional covariance matrix.
Noise- and outlier-robust test for jumps

- Theorem III forms the basis for a nonparametric noise- and outlier-robust test for jumps.
- Use suitably scaled measure of jumps by comparing $RV_N^*[Y]$ with either $BV_N^*[Y]$ or $QRV_N^*[Y]$, e.g.

$$N^{1/4} \left( RV_N^*[Y] - BV_N^*[Y] \right) \xrightarrow{d} N(0, 1) \cdot \sqrt{\Sigma_{11}^* + \Sigma_{22}^* - 2\Sigma_{12}^*}$$

- In practice, a transformation (using the delta method) can improve finite sample properties of the test, e.g., a ratio- or log-based version. We found that the log-based test performs well.
- Infeasible result, as $\Sigma^*$ is unknown!
Estimating $\Sigma^*$

- In order to construct a feasible jump test, we need to estimate the conditional covariance matrix $\Sigma^*$.
- We can construct estimators of the individual entries of $\Sigma^*$, for example

$$\hat{\Sigma}^{*}_{11} = \frac{N^{-1/2}}{\theta^2 \psi^2} \sum_{i=K}^{N-2K+1} |\bar{Y}^N_i|^2 \left( \sum_{l=-K+1}^{K-1} \left( |\bar{Y}^N_{i+l}|^2 - |\bar{Y}^N_{i+K}|^2 \right) \right) \xrightarrow{p} \Sigma^*_{11}$$

- Problem: The full estimated covariance matrix $\hat{\Sigma}^*$ often not positive semi-definite.
- We propose a positive semi-definite block subsample estimator of $\Sigma^*$, which has an intuitive form.
We restrict attention to the $2 \times 2$ submatrix of $\Sigma^*$ holding the covariance structure of $(RV^*_N[Y], BV^*_N[Y])$.

We choose two frequencies $d$ and $L$, such that $L >> K$ and $dL = o(N)$. Here $d =$ number of subsamples, $L =$ block length.

Let

$$RV^*_{N,m}[Y] = \frac{1}{K \psi_2^K} \sum_{i \in J_m} |\bar{Y}_i|^2 - \frac{\psi^{1}_K}{\theta^{2} \psi^{2}_K} \hat{\omega}_AC^2, \quad m = 1, \ldots, d$$

where

$$J_m = \{i : 0 \leq i \leq N - K + 1 \text{ and } (m - 1 + jd)L \leq i < (m + jd)L \text{ for some } j\}.$$
$RV_{N,1}^*[Y] \quad RV_{N,2}^*[Y] \quad \cdots \quad RV_{N,d}^*[Y]$

$\bar{Y}_i^N \quad \bar{Y}_i^N \quad \cdots \quad \bar{Y}_i^N \quad \bar{Y}_i^N \quad \bar{Y}_i^N \quad \cdots \quad \bar{Y}_i^N$

$\Delta_i^N Y$

Illustration of subsampler
Note that, asymptotically, $RV_{N,m}^*[Y]$ are mutually independent, because they are based on non-overlapping increments.

Moreover, they satisfy the same CLT as $RV_N^*[Y]$, but with convergence rate $N^{1/4}/\sqrt{d}$.

It is intuitive that a good proxy for $\Sigma_{11}^*$ is given by

$$\hat{\Sigma}_{11}^* = \frac{1}{d} \sum_{m=1}^{d} \left( \frac{N^{1/4}}{\sqrt{d}} (RV_{N,m}^*[Y] - IV) \right)^2$$

As the IV is unknown, we replace it with $RV_N^*[Y]$.

We then construct $BV_{N,m}^*[Y]$ in a similar fashion.
Finally, we set

\[
T_{N,m} = \frac{N^{1/4}}{\sqrt{d}} \left( RV^*_N,m[Y] - RV^*_N[Y], BV^*_N,m[Y] - BV^*_N[Y] \right)',
\]

and compute

\[
(\hat{\Sigma}^*_{ij})_{1 \leq i,j \leq 2} = \frac{1}{d} \sum_{m=1}^{d} T_{N,m} T'_{N,m} \xrightarrow{p} (\Sigma^*_{ij})_{1 \leq i,j \leq 2},
\]

The estimator is positive semi-definite by construction.

Unreported simulations show that \((\hat{\Sigma}^*_{ij})_{1 \leq i,j \leq 2}\) is largely unbiased if \(L\) is not too small. Also, it improves in an MSE sense by choosing larger values of \(d\).
Simulation study

- We simulate from a number of models, including models with stochastic volatility (1- or 2- factors), leverage, jumps and outliers.
- We use $N = 10,000$ and pollute $X$ using a noise ratio of $\gamma = 0.25$ (see, e.g., Oomen, 2006).
- Noise-robust pre-averaging estimators are computed using $\theta = \{0.10; 0.25; 0.50\}$.
- We base the QRV on absolute returns using $m = 3$ and $\lambda = 2/3$.
- This calibration is known as the MedRV (see, e.g., Andersen, Dobrev and Schaumburg, 2008).
### Table: Relative bias of pre-averaging estimators.

<table>
<thead>
<tr>
<th>model (down) // θ (right)</th>
<th>(RV_N^*[Y])</th>
<th>(BV_N^*[Y])</th>
<th>(BV_N^*<a href="%CF%84">Y</a>)</th>
<th>(MedRV_N^*<a href="%CF%84">Y</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
</tr>
<tr>
<td>SV</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 0.99</td>
</tr>
<tr>
<td>SV-LEV</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 0.99</td>
</tr>
<tr>
<td>SEV-ND</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 0.99</td>
</tr>
<tr>
<td>SV2F-LEV</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
</tr>
<tr>
<td>BMJ((n_J = 1, ν_J = \frac{1}{4}))</td>
<td>1.25 1.25 1.25</td>
<td>1.03 1.04 1.05</td>
<td>1.00 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
</tr>
<tr>
<td>BMJ((n_J = 5, ν_J = \frac{1}{4}))</td>
<td>1.25 1.25 1.25</td>
<td>1.05 1.08 1.10</td>
<td>1.00 1.01 1.03</td>
<td>1.00 1.01 1.02</td>
</tr>
<tr>
<td>BMJ((n_J = 10, ν_J = \frac{1}{4}))</td>
<td>1.25 1.25 1.25</td>
<td>1.07 1.10 1.13</td>
<td>1.01 1.03 1.06</td>
<td>1.01 1.02 1.05</td>
</tr>
<tr>
<td>BMJ((n_J = 5, ν_J = \frac{1}{2}))</td>
<td>1.50 1.50 1.50</td>
<td>1.08 1.12 1.16</td>
<td>1.00 1.01 1.02</td>
<td>1.00 1.01 1.02</td>
</tr>
<tr>
<td>BM-outlier</td>
<td>1.00 1.00 1.00</td>
<td>0.99 1.00 1.00</td>
<td>1.01 1.00 1.00</td>
<td>1.00 1.00 1.00</td>
</tr>
</tbody>
</table>

**Note.** This table reports the relative bias for the pre-averaging estimators \(RV_N^*[Y]\), \(BV_N^*[Y]\) and \(MedRV_N^*[Y]\). In the simulations, we set \(N = 10,000\) and \(γ = 0.25\). The \(MedRV_N^*[Y]\) is a special case of the QRV estimator based on absolute returns, using \(m = 3\) and \(λ = 2/3\). The bias measure is equal to 1 for an unbiased IV estimator. Pre-averaging estimators based on a threshold to pre-trim \((\bar{Y}_i^N)\) are denoted with \((τ)\).
As seen in Table 1, $BV^*_N[Y]$ is upward biased in the presence of jumps. This type of effect is also known from $BV_N[X]$.

The bias is also present in the jump-robust $QRV^*_N[Y]$, although to a slightly lesser extent (not reported).

In finite samples, this induces a downward bias in the estimated jump proportion and reduces the power of the jump test.

To alleviate the bias, we experiment with a threshold in the pre-averaged return series $\bar{Y}_i^N$.

The idea is related to the work of Aït-Sahalia and Jacod (2009), Corsi, Pirino and Renò (2010) or Mancini (2006), but there are some deviations in the workings of the threshold.
Note that under a Brownian motion with i.i.d. noise, the asymptotic distribution (as $N \to \infty$) of $\bar{Y}_N$ is given by

$$N^{1/4} \bar{Y}_N \xrightarrow{a} N \left( 0, \psi_2 \theta \sigma^2 + \psi_1 \frac{1}{\theta} \omega^2 \right)$$

where $\psi_1 = \lim_{K \to \infty} \psi_1^K$ and $\psi_2 = \lim_{K \to \infty} \psi_2^K$

Thus, we can set a threshold by computing

$$\tau = q_{1-\alpha} \times \sqrt{\psi_2^K \theta \sigma^2 + \psi_1^K \frac{1}{\theta} \omega^2} \times N^{-\varpi},$$

where $q_{1-\alpha}$ is an appropriate high quantile from the $N(0, 1)$ distribution and $\varpi \in (0, 0.25)$. 
Throughout, we work with $\alpha = 0.001$ and $\varpi = 0.20$, which produces good results in our simulations.

In practice, we also use plug-in estimators of unknown quantities, i.e. we make the replacements $\sigma^2 \to \hat{\sigma}^2$, $\omega^2 \to \hat{\omega}^2_{AC}$.

This amounts to a two-stage procedure, where $\hat{IV}$ and $\hat{\omega}^2$ are pre-estimated in order to set the threshold.

After filtering the data, we then re-compute the estimator.

As the original jump-robust estimator used to pre-estimate $\hat{IV}$ is slightly upward biased in the presence of jumps, the suggested procedure should also lead to conservative levels of $\tau$. 
A naive threshold simply throws away all extreme observations, i.e. pre-averaged returns which satisfy

$$|\bar{Y}_i^N| > \tau.$$
Nonetheless, the connection between \((\Delta_i^N Y)\) and \((\bar{Y}_i^N)\) can be exploited by searching and selectively discarding noisy returns.

Simple rule: If a breach of \(\tau\) is observed, we extract the raw noisy returns that are used to construct the pre-averaged returns in that sequence.

Then we discard the largest noisy return.

The procedure can probably be improved, but our simulations suggest that it does a reasonable job (see Table 1).
We illustrate the mechanics below.

Note. To the left is the noisy return series, $(\Delta_y^N Y)$, while the pre-averaged return series, $(\bar{Y}_y^N)$, is to the right. The threshold $\tau$ is the plotted with red lines. The orange area shows the part of $(\Delta_y^N Y)$ taken out for inspection, while the black circle highlights the discarded return.
Note. We report the simulated size of the feasible log-based jump test under model BM for $\theta = \{0.10; 0.25; 0.50\}$.
Note. We report the simulated size of the feasible log-based jump test under model SV2F-LEV for $\theta = \{0.10; 0.25; 0.50\}$. 
Power: model BMJ($n_J = 1, \nu_J = \frac{1}{4}$)

Note. We report the simulated size-adjusted power of the feasible log-based jump test under model BMJ($n_J = 1, \nu_J = \frac{1}{4}$) for $\theta = \{0.10; 0.25; 0.50\}$. 
Power: model BMJ\( (n_J = 5, v_J = \frac{1}{4}) \)

Note. We report the simulated size-adjusted power of the feasible log-based jump test under model BMJ\( (n_J = 5, v_J = \frac{1}{4}) \) for \( \theta = \{0.10; 0.25; 0.50\} \).
Note. We report the simulated size-adjusted power of the feasible log-based jump test under model BMJ($n_J = 5$, $v_J = \frac{1}{4}$) for $\theta = \{0.10; 0.25; 0.50\}$. 

Size without threshold: model SV2F-LEV
Power without threshold: model BMJ\((n_J = 1, \nu_J = \frac{1}{4})\)

**Note.** We report the simulated size-adjusted power of the feasible log-based jump test under model BMJ\((n_J = 5, \nu_J = \frac{1}{4})\) for \(\theta = \{0.10; 0.25; 0.50\}\).
Some remarks

- The test has good size and power under model BM, but it is over-sized under stochastic volatility. The problem gets smaller, when the pre-averaging parameter $\theta$ is increased.

- Increasing $\theta$, however, causes a slight drop in simulated power. Trade-off is in part influenced by setting a constant threshold.

- Nominal size is restored if we drop the threshold, but then the power of the test is eroded $\rightarrow$ Because upward bias in $BV^*_N[Y]$ and estimated standard errors deflates t-statistic.

- Practical compromise: Choose $\theta$ larger than theoretical minimum MSE choice would imply, but avoid excessive pre-averaging.
Data description

- We apply the pre-averaging technology to draw inference about jumps using a unique, extensive set of ultra high-frequency data.
- We extracted data from the NYSE TAQ database for the most recent configuration of DOW Jones (October, 2010), plus the two ETFs SPY and QQQQ.
- We analyze both transaction and quotation data (only results from transaction data are reported here).
- After “light” cleaning and aggregation, we are left with a total sample size of about 4.3 billion tick-by-tick observations.
Note. We plot the average annualized volatility of the noise- and outlier-robust estimators, averaged across the cross-section of stocks included in our empirical application, as a function of $\theta$. $RV_{5m}$ and $BV_{5m}$ are shown as a comparison.
Note. We plot the average estimated jump proportion, averaged across the cross-section of stocks included in our empirical application, as a function of $\theta$. The jump proportion estimated by using $RV_{5m}$ and $BV_{5m}$ is shown as a comparison.
**Regression analysis**

\[ BV^*_N[Y](\tau) = -0.075 + 0.996 \times RV^*_N[Y] \]

\[ BV_{5m}[Y] = 0.026 + 0.925 \times RV_{5m}[Y] \]

\(t_{b_0=0} = 0.831, t_{b_1=1} = -13.231\)

**Note.** The figure shows pairwise values of the average realised variance and bipower variation for each company in our selection of stocks (reported as a blue circle). We fit a regression line and test the hypothesis \(b_0 = 0\) and \(b_1 = 1\).
Jump test: Alcoa [AA]

Note. On the left is the noise- and outlier-robust jump test, while to the right is the low-frequency jump test based on 5-minute sampling. We also report the actual number of rejections based on the 5- and 1-% significance level.
Note. We plot an example of a real and simulated sample path, where a large intraday move in the price is observed over a short period of time. We then zoom into ultra high-frequency view of the sample path around the move.
Option trading with rehedging

- Our results point towards a less important role for jumps than reported in previous papers. Finding could, in part, be driven by infrequent sampling, microstructure noise and data resolution.

- To illustrate the economic importance of distinguishing between a burst in volatility versus real economic jumps in financial markets, we consider an example from option trading.

- We suppose an option trader sells a short-term at-the-money call option and hedges his position in the underlying (covered call strategy) → Initially, trader is delta neutral but short gamma.

- Due to transaction costs, the trader only rehedges his delta position after every 1% move in the underlying.
Simulation details

- We simulate from a scaled Brownian motion with no drift: \( X_t = \sigma W_t \).
  We assume the annualized volatility is 40%.
- We price the option with the Black-Scholes model. We assume the option has 1 day left to maturity and that the risk-free rate is zero.
- The initial stock price is 100 and the option is at-the-money, so strike is also 100.
- At a random position in the sample path, we place either a 2% jump in price or a 2% “burst in volatility” (cf. the simulated example path in previous figure).
As one would expect, the trader faced with jump risk has larger losses than the trader, which faces burst risk.

Short gamma traders face losses that are proportional to the square of the move in the underlying.

Thus, in contrast to a jump, a burst in volatility allows the trader the valuable opportunity of rehedging his position, as the underlying moves.
Note. The plot shows the distribution of the P&L for the option trader exercising the covered call strategy with rehedging. Also reported in the figure is the average loss to the trader expressed in percent of the premium.
Conclusion

- We formulate a model, where the “efficient” price is contaminated with noise and outliers. We show that pre-averaging alleviates both sources of bias.
- We also suggest a threshold elimination procedure and propose a positive semi-definite estimator of the asymptotic covariance matrix appearing in the CLT.
- A simulation study shows these estimators are good also in finite samples.
- Using an extensive set of ultra high-frequency data, we find a much lower jump proportion and much less jumps than previously reported.
Ideas for future work / improvements

- At current, we are using a fixed value of $\theta$, it is probably better to work with a data-driven choice. Not a simple problem! See, e.g., Hautsch and Podolskij (2010) for an MSE-based suggestion.
- Further refinement of the procedure suggested to do threshold estimation, e.g. to allow for time-varying threshold.
- Application to an OTC market, where there is no limit order book.
- Study the properties of pre-averaging, when there are potentially an infinite number of (small) jumps in the price process.