Separating Diffusive Returns from Jump Returns, Truncated Variance and Inference for IV

Guilherme Salomé

September 13, 2018

1 Setup and Review

We assume that log-prices evolve according to a jump-diffusion process:

$$dX_t = \sqrt{c_t} dW_t + J_t$$

where c_t is the variance process and J_t is a jump process (of finite activity). Notice that the process above does not have a drift term. This is because the drift is negligible at small time periods.

The process is sampled at equi-distant intervals:

$$X_{i\Delta_n} \text{ for } i = 0, 1, 2, \ldots, nT$$

where T is the number of days for which the process is observed, and n counts the number of observations within a day.

Remember that the process X has two parts: one that is continuous (diffusion), and one that is discontinuous (jump). Indeed:

$$\begin{aligned} X_t &= \underbrace{(X_t - \Delta X_t)}_{\text{continuous}} + \underbrace{\Delta X_t}_{\text{jumps}} \\ &= X_t^c + X_t^d \end{aligned}$$

Let's assume $T = 1$. Then the daily integrated variance is given by:

$$IV = \int_0^1 c_s ds$$

Given the geometric returns $r_i = \Delta_i^n X$ for $i = 1, 2, \ldots, n$, we studied two estimators for the integrated variance. The realized volatility (RV) and the bipower variance (BV):

$$\begin{aligned} RV &\equiv \sum_{i=1}^n r_i^2 \\ BV &\equiv \frac{\pi}{2} \sum_{i=2}^n |r_i r_{i-1}| \end{aligned}$$

RV and BV are estimators for IV, but RV is affected by the presence of jumps, while BV is jump robust. Next, we will discuss the theory for separating moves in the price (returns) that come from the continuous part of the process (X^c) and moves that come from the discontinuous part of the process (X^d).

2 Separating Diffusive Returns from Jump Returns

Let's denote the return at the i-th interval of day t by $r_{t,i}$:

$$\begin{aligned} r_{t,i} &\equiv X_{(i+(t-1)n)\Delta_n} - X_{(i-1+(t-1)n)\Delta_n} \\ &= \Delta^n_{i+(t-1)n} X \end{aligned}$$

for $t = 1, 2, \ldots, T$ and $i = 1, 2, \ldots, n$. To simplify the notation, let $t_i \equiv (i + (t-1)n)\Delta_n$. Then, the return on any interval is given by

$$\begin{aligned} r_{t,i} &= X_{t_i} - X_{t_{i-1}} \\ &= X_{t_i} - X_{t_i - \Delta_n} \end{aligned}$$

The idea is to classify each of the moves (returns) as continuous (diffusive) or discontinuous (jump):

$$r_{t,i} = r^c_{t,i} + r^d_{t,i}$$

Using the jump-diffusion process, we can write:

$$X_{t_i} = X_{t_i - \Delta_n} + \int_{t_i - \Delta_n}^{t_i} \sqrt{c_s} dW_s + \sum_{t_i - \Delta_n < s \le t_i} \Delta J^2_s$$

If there were no jumps:

$$\begin{aligned} r_{t,i} &= X_{t_i} - X_{t_i - \Delta_n} \\ &= \int_{t_i - \Delta_n}^{t_i} \sqrt{c_s} dW_s \\ &\approx \underbrace{\sqrt{\int_{t_i - \Delta_n}^{t_i} c_s ds}}_{\equiv \sigma_{t,i}} Z_{t,i} \\ &\overset{d}{\sim} \mathcal{N}\left(0, \sigma^2_{t,i}\right) \end{aligned}$$

If we could observe (we do not) the path of the variance process c_t on the interval $[t_i - \Delta_n, t_i]$, then we could compute $\sigma_{t,i}$. Now, the return over such a short time interval is approximately normal:

$$r_{t,i} \approx \sigma_{t,i} Z_{t,i}$$

The probability that $r_{t,i}$ is between 3 standard deviations of the mean is about 99.73%. The probability that $r_{t,i}$ is between 4 standard deviations of the mean is about 99.99%. That is, if there are no jumps in the time interval $[t_i - \Delta_n, t_i]$, then we expect most of the returns to fall within 4 standard deviations of the mean. However, we know jumps can occur at any time interval. If a jump occurs, then the magnitude of $r_{t,i}$ will be dominated by the jump and will far exceed 4 standard deviations of the mean.

This analysis motivates the following rule:

$$\begin{aligned} |r_{t,i}| &\le \alpha_n \sigma_{t,i} : \text{return is diffusive} \\ |r_{t,i}| &> \alpha_n \sigma_{t,i} : \text{return is jump} \end{aligned}$$

Where α_n is some threshold that decreases as $\Delta_n \to 0$. The use of a threshold to separate continuous returns from jump returns is due to Mancini (2001) and Mancini (2009), and was refined in Jacod and Protter (2012). The formal proof that the threshold correctly classifies diffusive moves from jump moves was developed in Li, Todorov, and Tauchen (2017).

The current practice is the following:

$$\begin{aligned} r_{t,i}^c &= r_{t,i} 1_{\left\{|r_{t,i}| \leq \alpha \Delta_n^{0.49} \sqrt{\tau_i BV_t}\right\}} \\ r_{t,i}^d &= r_{t,i} 1_{\left\{|r_{t,i}| > \alpha \Delta_n^{0.49} \sqrt{\tau_i BV_t}\right\}} \end{aligned}$$

where $\alpha \in [3.5, 4.5]$.

Let's understand the substitution of $\sigma_{t,i}$ by the term $\Delta_n^{0.49}\sqrt{\tau_i BV_t}$. First, we do not observe the path of c_t, so we cannot directly compute $\sigma_{t,i}$. However, we can estimate the total variation of a day (IV_t) using the bipower variance estimator (BV_t). But using the total variation leads to a problem: on a small interval the variation is not given by IV_t but only a fraction of it. This can be fixed by appropriately redistributing the total variation. Thus, instead of writing $r_{t,i} \approx \sigma_{t,i} Z_{t,i}$ we have:

$$\begin{aligned} r_{t,i} &\approx \sqrt{\Delta_n IV_t} Z_{t,i} \\ &= \Delta_n^{0.50} \sqrt{IV_t} Z_{t,i} \end{aligned}$$

When redistributing the total variance there is another problem that needs to be addressed: the variance process exhibits an intraday pattern. This intraday pattern refers to the fact that in the mornings the variance is higher than during lunch time, and after lunch time the variance increases again. To adjust for this effect, we scale IV_t depending on the time of the day, leading to:

$$r_{t,i} \approx \Delta_n^{0.50} \sqrt{\tau_i IV_t} Z_{t,i}$$

We call τ_i the diurnal pattern or the time-of-day factor.

Now, the local return $(r_{t,i})$ is approximately normally distributed:

$$r_{t,i} \overset{d}{\sim} \mathcal{N}\left(0, \Delta_n \tau_i IV_t\right)$$

To correctly separate the diffusive from the jump returns, we need a threshold that is slightly bigger than $\Delta_n^{0.50}\sqrt{\tau_i IV_t}$. Alternatively, we need a threshold that decreases slower than $\Delta_n^{0.50}\sqrt{\tau_i IV_t}$ when $\Delta_n \to 0$ to allow all diffusive moves to get through the threshold and exclude the jump moves. To achieve that we inflate the cutoff by:

$$\begin{aligned} \Delta_n^{-0.01} &= \left(\frac{1}{n}\right)^{-0.01} \\ &= n^{0.01} \\ &> 1 \end{aligned}$$

In summary, we separate returns using the cuttoff:

$$cut_{t,i} \equiv \alpha_n \Delta_n^{0.49} \sqrt{\tau_i BV_t}$$

The exponent for Δ_n does not necessarily have to be 0.49, but can actually be any number close to 0.50. The literature often writes this threshold as:

$$cut_{t,i} \equiv \alpha_n \Delta_n^{\varpi} \sqrt{\tau_i BV_t} \text{ for } 0 < a < \varpi < 0.50$$

where a is a lower bound determined by some other conditions (unimportant here).

3 Truncated Variance

Now that we can separate returns coming from the diffusive part of the model from the returns coming from the jump part, we can study a new estimator for the integrated variance.

The truncated variance is an estimator for the integrated variance that is robust to jumps. It is defined as:

$$\begin{aligned} TV_t &\equiv \sum_{i=1}^{n} (\Delta^n_{i+(t-1)n} X)^2 1_{\left\{\left|\Delta^n_{i+(t-1)n} X\right| \leq cut_{t,i}\right\}} \\ &= \sum_{i=1}^{n} (r^c_{t,i})^2 \end{aligned}$$

This estimator throws away moves with big jumps, retaining only the moves that are diffusive to estimate the variance.

Under regularity conditions, it is possible to show that:

$$\Delta_n^{-1/2}(TV_t - IV_t) \overset{d}{\approx} \mathcal{N}\left(0, 2\int_{t-1}^{t} c_s^2 ds\right)$$

Next, let's compare the truncated variance to the previous estimators.

4 Comparison of IV Estimators

The table below compares the asymptotic distribution of the IV estimators studied up to now:

Estimator	Definition	Rate of Convergence	Asymptotic Distribution
RV_t	$\sum_{i=1}^{n} r_{t_i}^2$	$\Delta_n^{-0.50}$	$\mathcal{N}\left(IV_t + \sum_{t_{i-1}<s\leq t_i} \Delta J_s^2, 2\int_0^1 c_s^2 ds\right)$
BV_t	$\sum_{i=2}^{n} \left\|r_{t_i} r_{t_{i-1}}\right\|^2$	$\Delta_n^{-0.50}$	$\mathcal{N}\left(IV_t, 2.61\int_0^1 c_s^2 ds\right)$
TV_t	$\sum_{i=1}^{n} (r_{t_i}^c)^2$	$\Delta_n^{-0.50}$	$\mathcal{N}\left(IV_t, 2\int_0^1 c_s^2 ds\right)$

Notice that RV is the only estimator that is affected by jumps. Under the presence of jumps RV estimates the integrated variance, but also adds in the squared jump returns. However, if there are no jumps, then the asymptotic variance of RV is smaller than that of BV. The TV estimator converges to the integrated variance whether there are jumps or not. Its asymptotic variance is smaller than that of BV, making it a more efficient jump robust estimator for the integrated variance.

5 Inference for IV

The theory indicates that for large n:

$$\Delta_n^{0.50}(TV_t - IV_t) \overset{d}{\sim} \mathcal{N}\left(0, 2\int_{t-1}^{t} c_s^2 ds\right)$$

In order to do inference for the integrated variance we need an estimator for its asymptotic variance.

5.1 QIV: Quartic Integrated Variation

Define:

$$QIV_t \equiv \int_{t-1}^{t} c_s^2 ds$$

QIV stands for Quartic Integrated Variation. It can be estimated by:

$$\widehat{QIV}_t \equiv (3\Delta_n)^{-1} \sum_{i=1}^{n} (r_{t,i}^c)^4$$

Let's verify the convergence and understand why the term $(3\Delta_n)^{-1}$ shows up. Remember that:

$$r_{t,i}^c = \sqrt{\int_{t_i-\Delta_n}^{t_i} c_s ds} Z_{t,i}$$

To simplify the argument, consider that c_s is constant:

$$r_{t,i}^c = \sqrt{\Delta_n c} Z_{t,i}$$

Then, let's study the convergence of just the sum of the continuous returns to the 4th power:

$$\begin{aligned} \sum_{i=1}^{n} (r_{t,i}^c)^4 &= \sum_{i=1}^{n} \Delta_n^2 c^2 Z_{t,i}^4 \\ &= (c^2 \Delta_n) \left(\Delta_n \sum_{i=1}^{n} Z_{t,i}^4 \right) \end{aligned}$$

We can use the law of large numbers on the right most term:

$$\Delta_n \sum_{i=1}^{n} Z_{t,i}^4 \to \mathbb{E}\Big[Z^4\Big] = 3$$

The left most term converges to 0:

$$c^2 \Delta_n \to 0$$

Putting both together we would get:

$$\sum_{i=1}^{n} (r_{t,i}^c)^4 \to 0$$

But what we want to estimate is c^2. In order to fix that, we add the term $3\Delta_n^{-1}$:

$$(3\Delta_n)^{-1} \sum_{i=1}^{n} (r_{t,i}^c)^4 = \frac{c^2}{3} \left(\Delta_n \sum_{i=1}^{n} Z_{t,i}^4 \right) \to c^2$$

These adjustments ensure that:

$$\widehat{QIV}_t \to \int_{t-1}^{t} c_s^2 ds$$

5.2 Confidence Interval for IV via Asymptotic Distribution

The estimator $\widehat{QIV}_t$ can be used to construct confidence intervals for IV_t. The asymptotic theory implies that for large n we have:

$$\Delta_n^{-\frac{1}{2}} \frac{TV_t - IV_t}{\sqrt{2\widehat{QIV}_t}} \overset{d}{\sim} \mathcal{N}(0,1)$$

We can use this result to construct a confidence interval for IV_t:

$$IV_t \in \left[TV_t - q_Z(\alpha/2)\sqrt{\Delta_n 2\widehat{QIV}_t}, TV_t + q_Z(\alpha/2)\sqrt{\Delta_n 2\widehat{QIV}_t}\right]$$

with probability $1-\alpha$, where $q_Z(\alpha/2)$ is the $\alpha/2$ quantile of the standard normal (e.g.: $q_Z(5\%/2) = -1.96$).

Notice that we can simplify the confidence interval by opening the expression inside the square-root:

$$\begin{aligned}\sqrt{\Delta_n 2\widehat{QIV}_t} &= \sqrt{2\Delta_n \frac{1}{3\Delta_n}\sum_{i=1}^{n}(r_{t,i}^c)^4} \\ &= \sqrt{\frac{2}{3}\sum_{i=1}^{n}(r_{t,i}^c)^4}\end{aligned}$$

Using this simplification we can write the confidence interval as:

$$CI(IV_t, \alpha) \equiv \left[TV_t - q_Z(\alpha/2)\sqrt{\frac{2}{3}\sum_{i=1}^{n}(r_{t,i}^c)^4},\ TV_t + q_Z(1-\alpha/2)\sqrt{\frac{2}{3}\sum_{i=1}^{n}(r_{t,i}^c)^4}\right]$$

5.3 Confidence Interval for IV via Bootstrapping

The bootstrap is an alternative way to create confidence intervals for a statistic of interest. We will discuss a bootstrapping scheme based on Hounyo (2013),Gonçalves and Meddahi (2009) and Dovonon et al. (2014).

The idea of the bootstrap is to use the data that we already have to obtain "new" samples from the same distribution. The idea is to draw random samples (with replacement) from the observed data. Then we use these new samples to compute the estimator of interest multiple times. We then use the fluctuations of the estimators to construct the confidence interval.

Let's fix $T = 1$, then we have n observations of continuous returns in a day:

$$r_1^c, r_2^c, \ldots, r_n^c$$

We want to draw new samples from these returns. However, the new samples must respect the heteroskedasticity (time-varying variance) over the day. To do so, we will partition the day into non-overlapping segments of length k_n, so that there are M divisions in a day.

$$\underbrace{r_1^c, r_2^c, \ldots, r_{k_n}^c}_{\text{1st}}, \underbrace{r_{k_n+1}^c, r_{k_n+2}^c, \ldots, r_{2k_n}^c}_{\text{2nd}}, \ldots, \underbrace{r_{(M-1)k_n+1}^c, r_{(M-1)k_n+2}^c, \ldots, r_{Mk_n}^c}_{\text{Mth}}$$

We must have $Mk_n \leq n$. In practice we will divide the day into an integer number of segments so that $Mk_n = n$.

Now, to generate a new random sample we draw k_n returns with replacement from each of the M divisions:

$$\underbrace{\tilde{r}^c_1, \tilde{r}^c_2, \ldots, \tilde{r}^c_{k_n}}_{\text{1st}}, \underbrace{\tilde{r}^c_{k_n+1}, \tilde{r}^c_{k_n+2}, \ldots, \tilde{r}^c_{2k_n}}_{\text{2nd}}, \ldots, \underbrace{\tilde{r}^c_{(M-1)k_n+1}, \tilde{r}^c_{(M-1)k_n+2}, \ldots, \tilde{r}^c_{Mk_n}}_{\text{Mth}}$$

Use this new random sample to compute the truncated variance estimator:

$$\widetilde{TV} = \sum_{i=1}^{Mk_n} (\tilde{r}^c_i)^2$$

Store $\widetilde{TV}$ and repeat the process: draw a new random sample from the original data and compute a new $\widetilde{TV}$ estimator. After repeating this a sufficient number of times (say 10000), we have a collection of estimators:

$$BS(TV) \equiv \left\{\widetilde{TV}_1, \widetilde{TV}_2, \ldots, \widetilde{TV}_{10000}\right\}$$

To construct the confidence interval for IV we compute the quantiles of the set above. For example, if we compute the 2.5% and 97.5% quantiles ($q_{BS(TV)}$) of the set, then the confidence interval is given by:

$$CI(TV, 5\%) \equiv \left[q_{BS(TV)}(2.5\%), q_{BS(TV)}(97.5\%)\right]$$

The bootstrap is useful because it avoids having to compute the asymptotic variance of the estimator, which can be quite hard in some cases. However, it does require a lot of computational power, since you need to draw a big number of random samples and re-compute the estimator every time, and then repeat the process for every day of your sample. It is important to know that the confidence intervals computed via bootstrap are no better or worse than the ones from the asymptotic theory. Additionally, there are stringent conditions to guarantee that the bootstrap confidence interval is valid. In this context, arguments from Li, Todorov, and Tauchen (2015) justify the use of the bootstrap.