Estimating the Local Variance

Let the dynamics of the log-price process be:

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

We have studied how to estimate the integrated variance from high-frequency data, and how to separate diffusive from jump returns. We will now use what we know about both topics to construct an estimator for the local variance.

We are interested in estimating the value of c_t (local variance) for some particular $t \in [0, T]$. For example, say we want the to know the value of the local variance at $t = 13:25:12$ November 20, 2017. With 5 minute data we cannot estimate the local variance down to the exact second, but we can estimate the local variance for the interval containing t:

$$t = 13:25:12 \in (13:25:00, 13:30:00]$$

To simplify the notation, let's assume $T = 1$. Let $i_t \in \mathbb{N}$ be such that:

$$t \in ((i_t - 1)\Delta_n, i_t \Delta_n]$$

We will divide the analysis in three parts. First, we will study the local variance estimator for when c_t is continuous and t is in the central part of the day. Second, we will study the local variance estimator for when c_t is continuous, but t is near the beginning or end of the day. Third, we will study the local variance estimator for when c_t is discontinuous.

1 Continuous Variance Process and Time in the Middle of the Day

Let's assume c_t is a continuous process and fix some time t we are interested in. The idea for estimating c_t is to construct a local version of the integrated variance estimator, where we sum up the squared continuous returns around the time t and divide by the size of the intervals:

$$\hat{c}_t \equiv \frac{1}{2k_n\Delta_n} \sum_{j=-k_n}^{k_n} (r^c_{i_t+j})^2$$

From the integrated variance theory we know that:

$$\sum_{j=-k_n}^{k_n} (r^c_{i_t+j})^2 \approx \int_{(i_t-k_n)\Delta_n}^{(i_t+k_n)\Delta_n} c_s ds$$

Using a change of variables to center c around time t, we get:

$$\begin{aligned}\sum_{j=-k_n}^{k_n} (r^c_{i_t+j})^2 &\approx \int_{(i_t-k_n)\Delta_n}^{(i_t+k_n)\Delta_n} c_s ds \\ &= \int_{-k_n\Delta_n+(i_t\Delta_n-t)}^{k_n\Delta_n+(i_t\Delta_n-t)} c(t+s)ds\end{aligned}$$

And because $i_t\Delta_n - t > 0$ is a small number when Δ_n is small, we can further approximate this integral by:

$$\begin{aligned}\sum_{j=-k_n}^{k_n} (r^c_{i_t+j})^2 &\approx \int_{-k_n\Delta_n}^{k_n\Delta_n} c(t+s)ds \\ \implies \hat{c}_t &\approx \frac{1}{2k_n\Delta_n}\int_{-k_n\Delta_n}^{k_n\Delta_n} c(t+s)ds\end{aligned}$$

Now, remember from calculus that for a differentiable function f:

$$\int_{-b}^{b} f(x+s)ds = F(x+b) - F(x-b)$$

where F is the antiderivative of f. The right hand side is something that looks like a derivative around x. In fact, if we divide both sides by $(x+b)-(x-b)$, we get:

$$\frac{1}{2b}\int_{-b}^{b} f(x+s)ds = \frac{F(x+b)-F(x-b)}{(x+b)-(x-b)}$$

Now, take limits with respect to b:

$$\begin{aligned}\lim_{b\searrow 0}\frac{1}{2b}\int_{-b}^{b} f(x+s)ds &= \lim_{b\searrow 0}\frac{F(x+b)-F(x-b)}{(x+b)-(x-b)} \\ &= F'(x) \\ &= f(x)\end{aligned}$$

That is, averaging the function f around the point x approximates the actual value of $f(x)$. We can use the result above to show that the local variance estimator converges to the value of c at time t. Indeed:

$$\begin{aligned}\hat{c}_t &\approx \frac{1}{2k_n\Delta_n}\int_{-k_n\Delta_n}^{k_n\Delta_n} c(t+s)ds \\ &\to c(t) \text{ as } k_n\Delta_n \to 0\end{aligned}$$

Due to the approximations there is a slight bias issue that can be corrected by multiplying a term (close to 1 if k_n is large) on the local variance estimator. Doing so leads to the following local variance estimator:

$$\hat{c}_t \equiv \frac{2k_n}{2k_n+1}\frac{1}{2k_n\Delta_n}\sum_{j=-k_n}^{k_n} (r^c_{i_t+j})^2$$

1.1 Choosing the Parameter k_n

The k_n parameter is the number of neighboring windows we use in estimating the local variance. How do we choose k_n? We need the estimator $\hat{c}_t$ to approximate the true value of c_t, in the sense that $\mathbb{E}[\hat{c}_t] \to c_t$. But we also want the estimator to be precise, so that $\text{Var}(\hat{c}_t) \to 0$.

Ait-Sahalia and Jacod (2014) show that for the procedure to work we need two things:

1. We need $k_n \to \infty$ so that we sum over a large number of terms to estimate the local integrated variance precisely;

2. But we need $k_n \Delta_n \to 0$ so that the window of integration shrinks to zero and hone in on c_t.

Ideally, we would want $k_n \to \infty$ very fast, so that the sum is over many terms and the integrated variance is estimated precisely (variance goes to zero). But we cannot have that, since we need to respect the condition $k_n \Delta_n \to 0$. The theory indicates that the optimal balance between the need to make $k_n \to \infty$ but not too fast is to have:

$$\begin{aligned} k_n &= K \Delta_n^{-\frac{1}{2}} \\ &= K\sqrt{n} \to \infty \text{ as } n \to \infty \\ k_n \Delta_n &= K \Delta_n^{\frac{1}{2}} \\ &= K \frac{1}{\sqrt{n}} \to 0 \text{ as } n \to \infty \end{aligned}$$

where K is a constant chosen by common sense. The condition above does not literally indicate that $k_n = \sqrt{n}$, but that k_n needs to be a lot smaller than n and somewhere around $\sqrt{n}$. The 5-minute data we are using implies that $\sqrt{n} = \sqrt{77} = 8.77$. Common sense suggest that we need to keep k_n between about 5 to 13, with sensitivity checks.

2 Continuous Variance Process and Time at Beginning/End of the Day

When t is close to $9:30$ or $16:00$ there might not be enough returns around time t to compute the local estimators. All we need to do is adjust the limits of the summation so that we do not run off the edge into the previous or next day, where the level of volatility might be totally different. Thus, we write the local variance estimator as:

$$\hat{c}_t \equiv \frac{1}{N_t \Delta_n} \sum_{j=-J_1}^{J_2} (r^c_{i_t+j})^2$$

where J_1 and J_2 are selected to ensure that $1 \leq i_t + j \leq n$, and N_t is the number of terms in the sum. If t is at the beginning of the day, then we will have $J_2 = k_n$, but J_1 equal to a smaller value. If t is at the end of the day, then we will have $J_1 = k_n$, but J_2 equal to a smaller value.

3 Discontinuous Variance Process

Consider c_t to be discontinuous at t. Let's define its left and right limits as:

$$c_t^- \equiv \lim_{\tau \nearrow t} c_\tau$$
$$c_t^+ \equiv \lim_{\tau \searrow t} c_\tau$$

Because of the discontinuity, we have $c_t^- \neq c_t^+$. We can have $c_t^- < c_t^+$ or $c_t^- > c_t^+$ since the variance can jump in either direction. Therefore, we can form estimators for the left and right limits of the variance process. To do so, we use the same idea as before, but instead of summing squared returns around time t, we sum squared returns before time t to estimate c_t^-, and squared returns after time t to estimate c_t^+.

$$\hat{c}_t^- \equiv \frac{1}{k_n \Delta_n} \sum_{j=1}^{k_n} (r_{i_t - j}^c)^2$$
$$\hat{c}_t^+ \equiv \frac{1}{k_n \Delta_n} \sum_{j=1}^{k_n} (r_{i_t + j}^c)^2$$

If the time t is at the end or the beginning of the day we need to adjust the number of terms being summed as before. The optimal balance between sending $k_n \to \infty$ but not too fast requires $k_n = K\sqrt{n}$.