

1 Notation

- 1-second last mid price was observed for 1-month
- Daily volatility is being estimated
- Sampling period : K ($60 * 10$) (FIXME:10 minute; why?)
- Parameter estimation period : M ($60 * 60 * 24 = 86400$) (FIXME: 1 day; why?)
- Sequence length : n ($86400 * 30 = 2592000$)
- time : $0 < t < n$
- log-price : x_t
- log-return : r_t

2 Jump adjust

Removes jumps and stitch the log-price

2.1 Truncation

Let truncation function $\Psi_\tau(x) = \text{sgn}(x) \min(|x|, \tau)$.
For the truncated process $\Psi_\tau(r_t)$, $-\tau \leq \Psi_\tau(r_t) \leq \tau$
From Hoeffding's lemma

$$E[\exp(\lambda r_t)] \leq \exp\left(\frac{\lambda^2 \tau^2}{2}\right)$$

$$\tau^2 < \sigma^2 \text{ ensures } E[\exp(\lambda r_t)] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right) \text{ (sub-Gaussian)}$$

So truncation was applied using,

$$\tilde{r}_t = \Psi_\tau(r_t) = \text{sgn}(r_t) \min(|r_t|, \tau = \sigma_t)$$

where $\sigma^2 = \frac{1}{M} \sum_{i=t}^{t+M} r_i$

2.2 Truncation with MLE

Huber loss of truncated series was minimized with L1 regularization, using daily volatility as prior.

- Average daily volatility : $\sigma^2 = \frac{1}{n-M} \sum_{t=1}^{n-M} \left(\frac{1}{M} \sum_{i=t}^{t+M} r_i \right)^2$

- Huber loss :

$$l_{\tau'}(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \leq \tau' \\ |x| - \frac{\tau'^2}{2} & \text{if } |x| > \tau' \end{cases}$$

$$\text{with } \tau' = \sigma\sqrt{M} \sim \sqrt{\sum_t^{t+M} r_t^2}$$

- Truncation : $\Psi_\tau(x) = \text{sgn}(x) \min(|x|, \tau)$
- Optimal truncation threshold $\hat{\tau} = \arg \min \left(l_{\tau'} \left(\sigma - \sqrt{\frac{1}{M} \sum_t^{t+M} \Psi(r_t)^2} \right) \right)$

2.3 Benchmark

Does not remove jump noise for benchmark purpose

3 Microstructure noise adjust

Removes high-frequency component

3.1 Pre-Averaging

Windowing

$$r_{PA} = r_t * w$$

was applied to the raw return r_t with window size = K and window function $w(s) = \min(s+1, (K-s))$

3.2 Fractional diffusion

Models price process as fractional diffusion, and removes drift.

Derivative with order $\neq 0.2$ is considered as drift, because stationarity is achieved (ADF test with $p=0.01$) at 0.2-th derivative. Fractional differentiation is an analytic continuation of discrete differentiation.

$$\frac{\partial^n x_t}{\partial t^n} = x(t) * w(s) \quad \text{where } w(s) = ({}_n C_s (-1)^s)$$

$$\frac{\partial^n x_t}{\partial t^n} = x(t) * w(s) \quad \text{where } w(s) = \left(\frac{\Gamma(n+1)}{\Gamma(s+1)\Gamma(n-s+1)} (-1)^s \right)$$

See : Jiahao Jiang, Bing Miao; A study of anomalous stochastic processes via generalizing fractional calculus. Chaos 1 February 2025; 35 (2): 023156. <https://doi.org/10.1063/5.0244009>

3.3 Padé transform

Models jumps as Lorentz peak of oscillator resonance.

4 Volatility estimation

Estimate volatility, for given raw log-price and denoised log-price.

4.1 RV

Baseline. $\sigma_t = \frac{1}{K} \sum_{i=t}^{t+K} r_i^2$

4.2 TSRV

Two-scale realized volatility

$$(\sigma_t^{TRV})^2 = \frac{1}{K} \sum_{i=t}^{t+K} (r_i^{PA})^2 - \frac{1}{2} \sum_{i=t}^{t+K} r_i^2 \sum_{s=1}^{K-1} (w(s) - w(s-1))^2$$

```
logprc = self.logprc.copy() x1 = np.square(logprc[:-K] - logprc[K:]).mean()  
x2 = np.square(np.diff(logprc)).mean() tsrv = x1 - x2
```

4.3 PRV

Pre-averaging realized volatility with noise adjustment

$$(\sigma_t^{PRV})^2 = \frac{1}{K} \sum_{i=t}^{t+K} (r_i^{PA})^2 - \frac{1}{2} \sum_{i=t}^{t+K} r_i^2 \sum_{s=1}^{K-1} (w(s) - w(s-1))^2$$

where w is window function.

5 Evaluation

Compare methods based on information measure

5.1 KLD of Conditional distribution

Measures the amount of information in time Filtration \mathcal{F}_t .

$$\text{KL}(P|P_0)$$

where $P = P(\sigma_t|\sigma_{t-1}) = P(\sigma_t|\mathcal{F}_t)P(\sigma_{t-1})$ and $P_0 = P(\sigma_t)P(\sigma_t) = P(\sigma_t)P(\sigma_{t-1})$.
P is estimated using kernel density estimation.

5.2 Self-similarity dimension

. Fractal dimension is used to measure information amount, using box-counting method.

5.3 Minkowski Measure

For given fractal dimension, measures persistency.