AARHUS UNIVERSITY

U- and V-Statistics for Semimartingales

Christian Schmidt, Aarhus University

Joint work with Mark Podolskij, Mathias Vetter, and Johanna F. Ziegel

The Setting

The Model

We consider a one-dimensional Itô-Semimartingale

$$X_t = x + \int_0^t a_s\, ds + \int_0^t \sigma_s\, dW_s + J_t$$

on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where a_s and σ_s are stochastic processes, W is a standard Brownian motion, and J_t is a jump-process.

The process is observed at discrete time points

$$X_{i/n}, \quad i = 0, \ldots, \lfloor nt \rfloor$$

and we are in the setting of high frequency data (i.e. $n \to \infty$). The increments of the process X are denoted by

$$\Delta_i^n X = X_{\frac{i}{n}} - X_{\frac{i-1}{n}}.$$

The Statistics

For any kernel function $H : \mathbb{R}^d \to \mathbb{R}$ we consider the U-statistic $U(H)_t^n$ of order d defined as

$$U(H)_t^n = \binom{n}{d}^{-1} \sum_{1 \le i_1 < \ldots < i_d \le \lfloor nt \rfloor} H(\sqrt{n}\Delta_{i_1}^n X, \ldots, \sqrt{n}\Delta_{i_d}^n X).$$

Similarly, the V-statistics $V_1(H)_t^n$ and $V_2(H)_t^n$ of order 2 are given by

$$V_1(H)_t^n = \sum_{1 \le i_1, i_2 \le \lfloor nt \rfloor} H(\Delta_{i_1}^n X, \Delta_{i_2}^n X),$$

$$V_2(H)_t^n = \frac{1}{n} \sum_{1 \le i_1, i_2 \le \lfloor nt \rfloor} H(\sqrt{n}\Delta_{i_1}^n X, \Delta_{i_2}^n X)$$

The main difference between the statistics is wether the scaling is inside or outside of the function H.

U-statistics

In the U-statistic case we consider $\mathbf{J_t = 0}$, i.e. X is a continuous semimartingale. We can show that we can replace the U-statistic $U(H)_t^n$ in the proofs of the asymptotic results by

$$\int_{\mathbb{R}^d} H(x_1, \ldots, x_d)\, F_n(dx_1, t) \ldots F_n(dx_d, t),$$

where F_n is an empirical process given by

$$F_n(x, t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \mathbb{I}_{\{\alpha_i^n < x\}}$$

and the $\alpha_i^n = \sigma_{(i-1)/n}\sqrt{n}\Delta_i^n W$ are approximations of the scaled increments $\sqrt{n}\Delta_i^n X$ of the process X.

Law of Large Numbers

In the notation from above we have convergence

$$F_n(x,t) \xrightarrow{\mathbb{P}} F(x,t) = \int_0^t \Phi_{\sigma_s}(x)\, ds,$$

where $\Phi_y(x)$ is the distribution function of the $\mathcal{N}(0, y^2)$ law.
Given that H is symmetric and of polynomial growth we obtain the law of large numbers

$$U(H)_t^n \xrightarrow{\mathbb{P}} U(H)_t = \int_{\mathbb{R}^d} H(x_1, \ldots, x_d) F(dx_1, t) \ldots F(dx_d, t).$$

Central Limit Theorem

For the central limit theorem we require some additional assumptions. The volatility process σ has to be a continuous Itô-semimartingale itself and the function H is assumed to be even in its arguments and continuously differentiable with H and H' of polynomial growth. Under the assumptions the sequence F_n of empirical processes fulfils the following stable CLT:

$$\sqrt{n}(F_n(x,t) - F(x,t)) \xrightarrow{st} \mathbb{G}(x,t),$$

where $\mathbb{G}$ is a process defined on an extension of $(\Omega, \mathcal{F}, \mathbb{P})$. The convergence is functional in t and in finite distribution sense in x. It can be shown that, conditionally on $\mathcal{F}$, $\mathbb{G}$ is Gaussian with mean zero and known covariance structure.

For the U-statistic itself we therefore obtain the CLT

$$\sqrt{n}(U(H)_t^n - U(H)_t) \xrightarrow{st} d \int_{\mathbb{R}^d} H(x_1, \ldots, x_d) \mathbb{G}(dx_1, t) F(dx_2, t) \ldots F(dx_d, t).$$

The limiting process is again conditionally Gaussian with mean zero and known variance. The variance can consistently be estimated by a slightly generalized U-statistic.

V-statistics

The limiting theory depends heavily on the form of the function H. We assume $H(x_1, x_2) = |x_1|^{p_1}|x_2|^{p_2} L(x_1, x_2)$ for some sufficiently smooth function L and $p_1, p_2 \in \mathbb{R}_+$.

Laws of Large Numbers

If H is continuous and under certain growth conditions on H and L, we find the following laws of large numbers:

$$V_1(H)_t^n \xrightarrow{\mathbb{P}} V_1(H)_t = \sum_{s_1, s_2 \le t} H(\Delta X_{s_1}, \Delta X_{s_2}) \qquad (p_1, p_2 > 2),$$

$$V_2(H)_t^n \xrightarrow{\mathbb{P}} V_2(H)_t = \sum_{s \le t} \int_0^t \tilde{H}(\sigma_u, \Delta X_s) du \qquad (p_1 < 2, p_2 > 2),$$

where $\tilde{H}(x_1, x_2) = \mathbb{E}[H(x_1 U, x_2)]$ with $U \sim \mathcal{N}(0,1)$.

Central Limit Theorems

We denote by $(T_n)_{n \in \mathbb{N}}$ a sequence of stopping times that exhausts the jumps of X. Further, on an extension of the original space, we define random variables $\kappa_n \sim \mathcal{U}([0,1])$, $\psi_{n\pm} \sim \mathcal{N}(0,1)$, all independent and independent of $\mathcal{F}$. Then let

$$R_n = \sqrt{\kappa_n}\sigma_{T_n-}\psi_{n-} + \sqrt{1-\kappa_n}\sigma_{T_n}\psi_{n+}$$

Under additional assumptions on X and differentiability conditions on H one can show

$$\sqrt{n}(V_1(H)_t^n - V_1(H)_t) \xrightarrow{st} \sum_{n_1, n_2 : T_{n_1}, T_{n_2} \le t} \sum_{i=1}^2 \partial_i H(\Delta X_{T_{n_1}}, \Delta X_{T_{n_2}}) R_{n_i} \qquad (p_1, p_2 > 3).$$

The $\mathcal{F}$-conditional law of the limit does not depend on the choice of the sequence (T_n). If σ and X do not jump at the same time, the limit is $\mathcal{F}$-conditionally Gaussian.

In the more complicated "mixed" case both jumps and the empirical process limit $\mathbb{G}$ appear.

$$\sqrt{n}(V_2(H)_t^n - V_2(H)_t) \xrightarrow{st} \sum_{n: T_n \le t} \int_0^t \tilde{\tilde{H}}(\sigma_u, \Delta X_{T_n}) R_n du + \sum_{s \le t} \int_{\mathbb{R}} H(x, \Delta X_s) \mathbb{G}(dx, t) \qquad (p_1 < 1, p_2 > 3),$$

where $\tilde{\tilde{H}}(x_1, x_2) = \mathbb{E}[\partial_2 H(x_1 U, x_2)]$ with $U \sim \mathcal{N}(0,1)$.

The limit is again $\mathcal{F}$-conditionally Gaussian if X and σ do not jump at the same time.

References

Podolskij, M., Schmidt, C., F. Ziegel, J.
Limit theorems for nondegenerate U-statistics of continuous semimartingales.
Annals of Applied Probability, **6**:2491–2526, 2014.

Jacod, J.
Asymptotic properties of realized power variations and related functionals of semimartingales.
Stoch. Proc. Appl., 118:517–559, 2008.