Estimation of the degree of jump activity, for irregularly sampled processes in presence of noise

Jean Jacod – Viktor Todorov

UPMC (Paris 6) – Northwestern University

The aim

Make inference on the jumps of a 1-dimensional process

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + \text{jumps driven by a Poisson measure}$$

observed at discrete times within the *fixed* time interval $[0,1]$,

Three features:

- The Brownian part $\int_0^t \sigma_s \, dW_s$ is typically dominating, but we are interested in jumps.
- The sampling times $0 = T(n,0) < T(n,1) < \cdots < T(n,i) < \cdots$ may be irregular, possibly random.
- There is a microstructure noise: instead of $X_{T(n,i)}$ we observe

$$Y_i^n = X_{T(n,i)} + \chi_i^n$$

(so tick-by-tick data, or all transactions data, can be used to do inference.

Notation

$$\Delta(n,i) = T(n,i) - T(n,i-1)$$

$$\Delta_i^n V = V_{T(n,i)} - V_{T(n,i-1)} \qquad V: \text{ any process}$$

$$\Delta V_t = V_t - V_{t-} \qquad V: \text{ any càdlàg process}$$

Regular sampling means $\Delta(n,i) = \Delta_n$.

Spot Lévy measures of X: the compensator ν of the jump measure of X is assumed to have the factorization

$$\nu(\omega, dt, dx) = dt\, F_{\omega,t}(dx)$$

(this is the "Itô semimartingale property" for the jumps). The measures $F_t = F_{\omega,t}$ are the spot Lévy measures, and $\int_0^t ds \int (x^2 \wedge 1)\, F_s(dx) < \infty$.

Warning: We do not want to "estimate" the jumps ΔX_t themselves, since

- the "big jumps" over $[0, 1]$ have no predictive value for the model.
- the "small" jumps are infinitely many, hence impossible to "estimate".

Rather, we will estimate the "degree of activity" of the jumps, and also the "intensity" of the small jumps (as specified below).

Assumptions on X (typically a log-price)

We have a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ on which:

$$X_t = X_0 + \int_0^t b_s\, ds + \int_0^t \sigma_s\, dW_s + \int_0^t \int_E \delta(s,z)(\underline{p} - \underline{q}\,)(ds,dz) + \int_0^t \int_E \delta'(s,z)\underline{p}\,(ds,dz)$$

$$\sigma_t = \sigma_0 + \int_0^t b_s^\sigma\, ds + \int_0^t H_s^\sigma\, dW_s' + \int_0^t \int_E \delta^\sigma(s,z)(\underline{p} - \underline{q}\,)(ds,dz) + \int_0^t \int_E \delta'^\sigma(s,z)\underline{p}\,(ds,dz)$$

- W and W' are two correlated Brownian motions.

- $\underline{p}$ is a Poisson measure on $\mathbb{R}_+ \times E$ with (deterministic) compensator $\underline{q}\,(dt,dz) = dt \otimes \eta(dz)$ (η is a σ-finite measure on the Polish space E).

- "standard" assumptions on the coefficients $b_t, b_t^\sigma, H_t^\sigma$: locally bounded, and (up to localization). b_t, H_t^σ satisfy for all finite stopping times $T \leq S$:

$$\mathbb{E}(\sup_{s\in[T,S]} |V_s - V_T|^2) \;\leq\; K\mathbb{E}(S-T) \tag{1}$$

- $|\delta(t,z)|^{r'}$, $1_{\{\delta'(t,z)\neq 0\}} \leq J(z)$ for some non-random η-integrable function J and some $r' \in [0,2)$, and the same for δ^σ, δ'^σ (up to localization).

MOREOVER, we need a structural assumption on the high-activity jumps of X, expressed in terms of the BG (Blumenthal-Getoor) index, or successive BG indices:

There is an integer $M \geq 0$, numbers $2 > \beta_1 > \cdots \beta_M > 0$, and nonnegative càdlàg processes $a_t^1, \ldots, a_t^M$, such that each $(a_t^m)^{1/\beta_m}$ satisfies (2), and the symmetrized Lévy measures $\breve{F}_t(A) = F_t(A) + F_t(-A)$ are such that the (signed) measure

$$F_t'(dx) = \breve{F}_t(dx) - \sum_{m=1}^{M} \frac{\beta_m a_t^m}{|x|^{1+\beta_m}} 1_{\{0<|x|\leq 1\}} \, dx$$

satisfies ($|F_t'|$ being the "absolute value" of F_t'):

$$|F_t'|([-x,x]^c) \leq \frac{\Gamma}{x^r} \qquad \forall x > \in (0,1].$$

Example:

$$X_t = X_0 + \int_0^t b_s \, ds + \int_0^t \sigma_s \, dW_s + \sum_{m=1}^{M} \int_0^t \gamma_{s-}^m \, dY_s^m + \int_0^t \int_E \delta'(s,z) \underline{p}(ds,dz)$$

with Y^m independent stable or tempered stable processes (with arbitrary dependencies with $\underline{p}$, and indices β_m) and γ^m's are Itô semimartingales.

We then have $a_t^m = |\gamma_t^m|^{\beta_m}$ (up to a multiplicative constant).

REGULAR OBSERVATIONS – NO NOISE

Choose a sequence k_n of integer ($\to \infty$) and set

$$L(y)_j^n = \frac{1}{k_n} \sum_{l=0}^{k_n-1} \cos(y(\Delta_{i+1+2l}^n X - \Delta_{i+2+2l}^n X)/\sqrt{\Delta_n}\,)$$

We take differences of two successive returns to “symmetrize” the jump measure around zero and kill the drift.

We have approximately, with $\chi(\beta) = \int_0^\infty \frac{\sin y}{y^\beta}\, dy$:

$$\mathbb{E}(L(y)_i^n \mid \mathcal{F}_i) \approx \exp\left(-y^2 \sigma_{i\Delta_n}^2 - 2 \sum_{m=1}^{M} \chi(\beta_m) y^{\beta_m} \Delta_n^{1-\beta_m/2} a_{i\Delta_n}^m\right)$$

hence the following is a estimator of the spot (squared) volatility $c_t = \sigma_t^2$ for $t \approx i\Delta_n$:

$$\widehat{c}(y)_j^n = -\frac{1}{y^2}\, \log\left(L(y)_j^n \bigvee \frac{1}{\log(1/\Delta_n)}\right)$$

Introducing a de-biasing term, we set

$$\widehat{C}(y)^n_t = 2k_n\Delta_n \sum_{j=0}^{[t/v_n]-1} \Big(\widehat{c}(y)^n_j - \frac{1}{y^2 k_n}\big(\sinh(y^2\widehat{c}(y)^n_j)\big)^2\Big),$$

where $\sinh(x) = \frac{1}{2}(e^x - e^{-x})$ is the hyperbolic sine. This "estimates"

$$C_t + \sum_{m=1}^{M} A^{m,n}(y)_t, \qquad \text{where}$$

$$A^{m,n}(y)_t = y^{\beta_m - 2}\Delta_n^{1-\beta_m/2} A^m_t, \qquad A^m_t = 2\chi(\beta_m)\int_0^t a^m_s\, ds$$

and the normalized error term is

$$Z(y)^n_t = \frac{1}{\sqrt{\Delta_n}}\Big(\widehat{C}(y)^n_t - C_t - \sum_{m=1}^{M} A^{m,n}(y)_t\Big).$$

The CLT: Let $\mathcal{Y}$ be a finite subset of $(0,\infty)$. Choose k_n and u_n such that

$$k_n\sqrt{\Delta_n} \to 0, \qquad k_n\Delta_n^{1/2-\varepsilon} \to \infty \ \ \forall\varepsilon>0, \qquad u_n \to 0, \qquad \frac{k_n\sqrt{\Delta_n}}{u_n^2} \to 0.$$

Theorem *We have the (functional) stable convergence in law:*

$$\Big(Z(u_n)^n,\ \big(\frac{1}{u_n^2}(Z(yu_n)^n - Z(u_n)^n)\big)_{y\in\mathcal{Y}}\Big) \overset{\mathcal{L}-s}{\Longrightarrow} \big(Z, ((y^2-1)\overline{Z})_{y\in\mathcal{Y}}\big),$$

where the limit is defined on an extension $(\widetilde{\Omega},\widetilde{\mathcal{F}},(\widetilde{\mathcal{F}}_t)_{t\geq 0},\widetilde{\mathbb{P}})$ of the original space $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\geq 0},\mathbb{P})$ and can be written as

$$Z_t = 2\int_0^t c_s\, dW_s^{(1)}, \qquad \overline{Z}_t = \frac{2}{\sqrt{3}}\int_0^t c_s^2\, dW_s^{(2)}.$$

where $W^{(1)}$ and $W^{(2)}$ are two independent Brownian motions, independent of the σ-field $\mathcal{F}$.

Estimation of β_1 (when $M=1$ for simplicity): Observe that

$$\begin{aligned}\overline{C}(y)^n_t &= \widehat{C}(u_n y)^n_t - \widehat{C}(u_n)_t \\ &= A^{m,n}(u_n y) - A^{m,n}(u_n)_t + \sqrt{\Delta_n}\,(Z(yu_n)^n_t - Z(u_n)_t)) \\ &= u_n^{\beta_1-2}(y^{\beta_1-2}-1)\Delta_n^{1-\beta_1/2}A^1_t + \sqrt{\Delta_n}\,(Z(yu_n)^n_t - Z(u_n)_t))\end{aligned}$$

The function $f(x) = \frac{4^x-1}{2^x-1}$ is C^∞ on $(1,2)$, with a C^∞ reciprocal function f^{-1}. Then a natural estimator for β_1 is, for example,

$$\widehat{\beta}^n_t = f^{-1}\Big(\frac{\overline{C}(4)^n_t}{\overline{C}(2)^n_t}\Big)$$

and we have

$$\frac{u_n^{\beta_1-2}}{\Delta_n^{(\beta_1-1)/2}}\,\big(\widehat{\beta}^n_t - \beta_1\big) \xrightarrow{\mathcal{L}-\mathrm{s}} Y$$

for some mixed centered normal y with a known (conditional) variance. The rate is thus almost $1/\Delta_n^{(\beta_1-1)/2}$, better than rates obtained by other methods when $\beta_1 > 4/3$.

The choice $k_n \approx 1/\sqrt{\Delta_n}$ and u_n going to 0 very slowly is in fact not "optimal" for this problem, so better (but complicated...) rates can actually be achieved.

One can also estimate $A_t^1 = 2\chi(\beta_1)\int_0^t a_s^1\,ds$, using for example

$$\widehat{A}_t^{n,1} = \overline{C}(2)_t^n\, \frac{1}{u_n^{\widehat{\beta}_1^n} - 2)(2^{\widehat{\beta}_1^n - 2} - 1)\Delta_n^{1-\widehat{\beta}_1^n/2}}$$

(on similarly estimate $\int_0^t a_s^1\,ds$).

Then we have a CLT for $\widehat{A}_t^{n,1}$, with the same kind of limit, and the same rate divided by $\log((1/\Delta_n)$ to an appropriate power.

NOISE AND IRREGULAR SAMPLING

Assumptions on the observation scheme

Assumption: The inter-observations lags are

$$\Delta(n, i+1) = \Delta_n \lambda_{T(n,i)} \Phi^n_{i+1}$$

- λ_t positive càdlàg adapted, satisfying

$$\mathbb{E}(|\lambda_T - \lambda_S|) \leq \mathbb{E}(T - S)$$

for all stopping times $S \leq T$ and $1/K \leq \lambda_t \leq K$ (up to localization).

- The variables Φ^n_i are positive, $\mathbb{E}(\Phi^n_i) = 1$, and $\sup_{n,i} \mathbb{E}(\Phi^n_i)^p) < \infty$ for all $p > 0$.
- The variables $(\Phi^n_i : i \geq 1)$ are mutually independent and independent of $\mathcal{F}_\infty$

  This implies in particular that $N^n_t = \sum_{i \geq 1} 1_{\{T(n,i) \leq t\}}$ satisfies

$$\Delta_n N^n_t \overset{\text{u.c.p.}}{\Longrightarrow} \Lambda_t := \int_0^t \frac{1}{\lambda_s} \, ds.$$

We denote by $\mathcal{H}^n_\infty$ the σ-field generated by $\mathcal{F}_\infty$ and all $\Phi^n_i : i \geq 1)$.

Two assumptions on the noise

We observe $Y_i^n = X_{T(n,i)} + \gamma'_{T(n,i)}\varepsilon_i^n$, where

(N1): The process γ' is an Itô semimartingale, and the variables $(\varepsilon_i^n : i \geq 1)$ are i.i.d., independent of $\mathcal{H}_\infty^n$, with

$$\mathbb{E}(\varepsilon_i^n) = 0, \qquad \mathbb{E}((\varepsilon_i^n)^2) = 1, \qquad \mathbb{E}(|\varepsilon_i^n|^p) < \infty.$$

(N2): The variables $(\varepsilon_i^n : i \geq 1)$ are independent, conditionally on $\mathcal{H}_\infty^n$ with

$$\mathbb{E}((\varepsilon_i^n)^p \mid \mathcal{H}_\infty^n) = \gamma_{T(n,i)}^{(p)}, \qquad \mathbb{P}(\varepsilon_i^n \in B \mid \mathcal{H}_\infty^n) = \mathbb{P}(\varepsilon_i^n \in B \mid \mathcal{H}_{T(n,i)}^n).$$

(where $(\mathcal{H}_t^n)$ is the smallest filtration containing $(\mathcal{F}_t)$ and for which all $T(n,i)$ are stopping times. Moreover, $\gamma_t^{(1)} = 0$ and $\gamma_t^{(2)} = 1$; moreover γ'_t and all $\gamma_t^{(p)}$ satisfy (2) and are $(\mathcal{F}_t)$-adapted.

(N1) is much stronger than (N2), and not very realistic. Later on, $\gamma_t = (\gamma'_t)^2$ (this is the "variance" of the noise).

An important example satisfying (N2) but not (N1).

Let $\rho_t^j \geq 0$ be càdlàg adapted nonnegative with $\sum_{j\in\mathbb{Z}} \rho_t^j = 1$ and $\rho_t^j = \rho_t^{-j}$ and $\sup_t \sum_{j\in\mathbb{Z}} \rho_t^j \, |j|^p < \infty$. For each n let $(Z_i^n : i \geq 1)$ be i.i.d. conditionally on $\mathcal{H}_\infty^n$, with density $x \mapsto \sum_{j\in\mathbb{Z}} \rho_{T(n,j)}^j 1_{[j,j+1)}(x)$. The observation at time $T(n,i)$ is

$$Y_i^n = [X_{T(n,i)} + Z_i^n],$$

so we have a additive (modulated) white noise *plus rounding.*

Remark: If we have "pure rounding", i.e. if we observe $[X_{T(n,i)}]$ (or $[X_{T(n,i)}] + \frac{1}{2}$ to "center" the noise), then no consistent estimator for C_t exists.

Pre-averaging

We choose 3 tuning parameters u_n, h_n, k_n all going to ∞: here $u_n > 0$ will be the argument of the empirical characteristic function below, and h_n, k_n (two integers) are window sizes.

The de-noising method is pre-averaging, but other methods could probably be used as well. Take a weight (or, kernel) function g on $\mathbb{R}$ with

$$\begin{array}{l} g \text{ is continuous, piecewise } C^1 \text{ with a piecewise Lipschitz derivative } g', \\ s \notin (0,1) \;\Rightarrow\; g(s) = 0, \qquad \int_0^1 g(s)^2 ds \;>\; 0, \end{array}$$

for example $g(x) = (x \wedge (1-x))\, 1_{[0,1]}(x)$. With a sequence $h_n \to \infty$ of integers, set

$$\begin{array}{lll} g_i^n = g(i/h_n), & \overline{g}_i^n \;=\; g_{i+1}^n - g_i^n & \\ \phi_n = \frac{1}{h_n} \sum_{i\in\mathbb{Z}} (g_i^n)^2, & \overline{\phi}_n = h_n \sum_{i\in\mathbb{Z}} (\overline{g}_i^n)^2, & \widetilde{\phi}_n^{(\beta)} = \frac{1}{h_n} \sum_{i\in\mathbb{Z}} |g_i^n|^\beta, \\ \phi = \int g(u)^2\, du, & \overline{\phi} = \int g'(u)^2\, du, & \widetilde{\phi}^{(\beta)} = \int |g(u)|^\beta\, du \end{array}$$

The *pre-averaged returns* of the observed values Y_i^n are

$$\widetilde{Y}_i^n = \sum_{j=1}^{h_n-1} g_j^n\, (Y_{i+j}^n - Y_{i+j-1}^n) = -\sum_{j=0}^{h_n-1} \overline{g}_j^n\, Y_{i+j}^n.$$

Initial estimators

For any $y > 0$, set

$$L(y)_i^n = \frac{1}{k_n} \sum_{l=0}^{k_n-1} \cos\big(yu_n(\widetilde{Y}^n_{i+2lh_n} - \widetilde{Y}^n_{i+(2l+1)h_n})\big)$$

(a proxy for the real part of the empirical characteristic function of the returns, over a window of $2h_nk_n$ successive observations). Taking a difference above allows us to "symmetrize" the problem.

Then, for any $y > 0$, a natural estimator for the "integrated volatility" over the time interval $[T(n,i), T(n,i+2h_nk_n)]$ is

$$\frac{2k_n}{y^2u_n^2\phi_n}\, v(y)_i^n, \qquad v(y)_i^n = -\log\Big(L(y)_i^n \bigvee \frac{1}{h_n}\Big).$$

We need to de-bias these estimators, to account for the noise, and also for some intrinsic distortion present even when there is no noise. With $f(x,y) = \frac{1}{2}\big(e^{2x-y} + e^{2x} - 2\big)$, set

$$\widehat{C}(y)_t^n = \tfrac{k_n}{y^2 u_n^2 \phi_n} \sum_{j=0}^{[N_t^n/2h_nk_n]-1} \Big(2v(y)_{2jh_nk_n}^n - \tfrac{1}{k_n}\, f(v(y)_{2jh_nk_n}^n, v(2y)_{2jh_nk_n}^n) \\ -\overline{\phi}_n y^2 u_n^2\, \textstyle\sum_{l=1} (\Delta_{2jh_nk_n+l}^n)^2\Big).$$

Finally, we set

$$Z(y)_t^n = \widehat{C}(y)_t^n - C_t - \frac{2}{\phi_n} \sum_{m=1}^M |y|^{\beta_m-2} u_n^{\beta_m-2} \widetilde{\phi}_n^{\beta_m} A_t^m$$

The basic CLT

Below, $\mathcal{Y}$ is any finite subset of $(0,\infty)$.

Theorem *under appropriate conditions on u_n, h_n, k_n, plus*

$$\frac{u_n^{\beta_1} h_n^3 \Delta_n}{u_n^{\beta_1} h_n^3 \Delta_n + u_n^4 (1 + h_n^2 \Delta_n)^2} \to \eta, \qquad \frac{h^2 \Delta_n}{1 + h_n^2 \Delta_n} \to \eta'$$

and with

$$v_n = k_n \sqrt{\frac{h_n^3 \Delta_n}{u_n^4 (1 + h_n^2 \Delta_n)^2 + u_n^{\beta_1} h_n^3 \Delta_n}}$$

for any $t > 0$ the variables $\big(v_n Z(y)_t^n\big)_{y\in\mathcal{Y}}$ converge $\mathcal{F}_\infty$-stably in law to $(Z(y)_t)_{y\in\mathcal{Y}}$, which is defined on an extension $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ and is, conditionally on $\mathcal{F}$, centered Gaussian with variance-covariance given by

$$\begin{aligned}&\widetilde{\mathbb{E}}(Z(y)_t Z(y')_t \mid \mathcal{F})\\ &= \int_0^t \big(\eta\psi(\beta_1, y, y') a_s^1 \lambda_s + (1-\eta) y^2 y'^2 \big(\eta' \phi \sigma_s^2 \lambda_s + (1-\eta')\overline{\phi}\gamma_s\big)^2 \frac{1}{\lambda_s}\, ds\end{aligned}$$

Estimation of β_1

Assume again $M = 1$. We use estimator analogous to the estimators in the regular no-noise case, that is

$$\widehat{\beta}_t^n = f^{-1}\Big(\frac{\overline{C}(4)_t^n}{\overline{C}(2)_t^n}\Big).$$

where $f(x) = \frac{4^x - 1}{2^x - 1}$ and

$$\begin{array}{rcl} \overline{C}(y)_t^n & = & \widehat{C}(y)_t^n - \widehat{C}(1)_t \\ & = & \frac{2}{\phi_n}\Big(u_n^{\beta_1 - 2}(y^{\beta_1 - 2} - 1)\Delta_n^{1-\beta_1/2} A_t^1 + (Z(y)_t^n - Z(1)_t))\Big) \end{array}$$

and we have (when the basic CLT holds):

$$u_n^{\beta_1/2}\big(\widehat{\beta}_t^n - \beta_1\big) \xrightarrow{\mathcal{L}-\mathrm{s}} Y$$

for some mixed centered normal Y with a known (conditional) variance. We similarly have estimators with an analogous CLT for A_t^1.

Problem: Choose $u_n, h_n, k_n)$ with the appropriate conditions, plus u_n as large as possible.

Reporting only the rate for β_1, we find the following (sub)-optimal rates (depending on the value of β_1, and on an arbitrary $\varepsilon > 0$).

- If $\sigma_t \equiv 0$ and (N1): $$\begin{array}{ll} 1/\Delta_n^{\frac{\beta_1}{10-4\beta_1}(1-\varepsilon)} & \text{if } \beta_1 \leq \frac{3}{4} \\ 1/\Delta_n^{\frac{\beta_1}{7}(1-\varepsilon)} & \text{if } \frac{3}{4} \leq \beta_1 \leq \frac{3}{2} \\ 1/\Delta_n^{\frac{\beta_1}{4+2\beta_1}(1-\varepsilon)} & \text{if } \beta_1 \geq \frac{3}{2} \end{array}$$

- If $\sigma_t \equiv 0$ and (N2): $$\begin{array}{ll} 1/\Delta_n^{\frac{3\beta_1}{30-11\beta_1}(1-\varepsilon)} & \text{if } \beta_1 \leq \frac{3}{4} \\ 1/\Delta_n^{\frac{3\beta_1}{21+\beta_1}(1-\varepsilon)} & \text{if } \frac{3}{4} \leq \beta_1 \leq \frac{3}{2} \\ 1/\Delta_n^{\frac{3\beta_1}{12+7\beta_1}(1-\varepsilon)} & \text{if } \beta_1 \geq \frac{3}{2} \end{array}$$

- If σ_t not 0 and (N1): $\begin{array}{ll} 1/\Delta_n^{\frac{\beta_1}{16-5\beta_1}(1-\varepsilon)} & \text{if } \beta_1 \leq \frac{16}{11} \\ 1/\Delta_n^{\frac{3\beta_1}{32-4\beta_1}(1-\varepsilon)} & \text{if } \beta_1 \geq \frac{16}{11} \end{array}$

- If σ_t not 0 and (N2): $1/\Delta_n^{\frac{3\beta_1}{48-11\beta_1}(1-\varepsilon)}$

Some problems:

1 - Optimality concerning β_1

2 - When $M \geq 2$, what does the rate for β_1 become ?, and can we estimate $\beta_2, \beta_3, \ldots$?

3 - How to choose in practice $(u_n h_n, k_n)$ (so far, only “mathematical” results are known.