Statistics for high-frequency observations of a stochastic process

Jean Jacod

Université Pierre et Marie Curie (Paris 6)

The aim

We have a stochastic process X, one-dimensional for simplicity. about which we want to do some kind of statistical inference. We are in the following setting:

- A finite horizon: X is indexed by time, say over the interval $[0, 1]$.

- A single path is observed (we cannot repeat the same "experiment")

- Discrete observations: we only observe the values $X_{i/n}$ for $i = 0, 1, \ldots, n$ (we only consider here "regular sampling", but irregular sampling could also be considered).

- Possibly noisy observations: We mainly consider the case where $X_{i/n}$ is exactly observed for $i = 0, 1, \ldots, n$, but in many practical examples one really observe $X_{i/n} + \varepsilon_i$, where ε_i is an observation noise.

- High-frequency setting: no consistent inference whatsoever can be done in the above setting when n is fixed. So, we consider the situation where n is "large", meaning looking at asymptotic inference by letting $n \to \infty$.

Examples of high-frequency data are more and more common, in biology, telecommunications, or finance, for example. For concreteness we focus here on financial data: X_t represents the price (or more often the log-price) of an asset at time t. Then

- A finite horizon: X can be observed over a very long period of time, but it is certainly not stationary. Its characteristics change with time, so for inference we need to restrict our attention to (relatively) short periods in which those characteristics are “resonably” constant.

- A single path: There are many other assets, but their prices have different characteristics than X; so, impossible to observe several paths of the same process X.

- Discrete observations: The price is really observed (or known) when transactions (or perhaps changes of quotes) happen.

- Possibly noisy observations: Typically X is modeled as a continuous process, or a discontinuous one with non trivial continuous part. However, prices are recorded as a multiple of a basic unit, say 1 cent. So there is a rounding error, and in finance there is also the so-called “microstructure noise”.

- High-frequency setting: the time interval $[0, 1]$ is perhaps one day, or one month. For liquid stocks there are typically several transactions within each second, meaning that during a working day we have more than $50,000$ observations.

What can be estimated ?

A very simple example (the Black-Scholes model): The log-price has the form

$$X_t = bt + \sigma W_t,$$

where $b \in \mathbb{R}$ is the drift, $\sigma > 0$ the square-root of the volatility, and W is a standard Brownian motion.

1) If $t \mapsto X_t$ is fully observed on $[0,t]$: if $\Delta_i^n X = X_{i/n} - X_{(i-1)/}$, then $\sum_{i=1}^n (\Delta_i^n X)^2 \xrightarrow{\mathbb{P}} \sigma^2$. Then, if X is fully observed, one knows in principle σ exactly, up to a null set (i.e., the probability of an error is 0).

On the other hand, b cannot be determined for sure, because the laws of X (over $[0,1]$) corresponding to any value of b are all equivalent (of course, in this case there is an optimal estimator of b, which is X_1, and the estimation error is $\mathcal{N}(0,\sigma^2)$).

Note also that if we observe X over $[0,\infty)$, in contrast, X_t/t converges to b as $t \to \infty$, so b is also known in this case. But this is due to the stationarity of (the increments) of X, a property that we want to avoid.

2) If $t \mapsto X_t$ is discretely observed: Again no consistent estimators for b, of course.

For σ, the convergence $\widehat{C}^n := \sum_{i=1}^n (\Delta_i^n X)^2$ to σ^2 has the rate $\sqrt{n}$, and in fact $\sqrt{n}\,(\widehat{C}^n - \sigma^2)$ converges in law to $\mathcal{N}(0, 2\sigma^4)$, as $n \to \infty$.

This is elementary, since the variables $\Delta_i^n X$ are i.i.d. as i varies, with law $\mathcal{N}(b/n, \sigma^2/n)$: a standard parametric problem. The MLE for σ^2 is not $\widehat{C}^n$ but $\sum_{i=1}^n (\Delta_i^n X - X_1/n)^2$, but its asymptotic behavior is the same as for $\widehat{C}^n$.

What can be estimated - 2 ?

The continuous Itô semimartingale case:

$$X_t = X_0 + \int_0^t b_s\,ds + \int_0^t \sigma_s\,dW_s$$

(b_t and $\sigma_t > 0$ are now processes themselves).

- As before, no way for estimating b_t.

- As before, if the whole path $t \mapsto X_t$ is observed, one knows the process σ_t on $[0,1]$. Indeed,

$$\widehat{C}_t^n = \sum_{i=1}^{[nt]} (\Delta_i^n X)^2 \xrightarrow{\mathbb{P}} C_t = \int_0^t \sigma_s^2\,de$$

(convergence in probability), where C_t is the quadratic variation of X. So in principle $(C_t)_{t\in[0,1]}$ is known, hence $(\sigma_t)_{t\in[0,1]}$ as well (up to null sets...)

Our aim becomes: “estimate” C_t for all t, which are the only quantities (besides X_t itself) which can be retrieved from the observations.

This is NOT a standard problem, since C_t is random. However one can prove:

$$\sqrt{n}\,\big(C_t^n - C_t\big) \text{ converges stably in law to } \sqrt{2}\int_0^t \sigma_s^2\,dB_t,$$

where B is another Brownian motion, independent of X. Equivalently, the limit is $N\sqrt{Q_t}$, where $Q_t = 2\int_0^t \sigma_s^4\,ds$ is called “quarticity” and N is $\mathcal{N}(0,1)$ independent of Q_t: the limit is “mixed normal” and the stable convergence (stronger than mere convergence in law) was introduced by Renyi (1963) in an analogous statistical context.

Moreover, one can also prove that

$$\widehat{Q}_t^n = \frac{2n}{3}\sum_{i=1}^{[nt]}(\Delta_i^n X)^4 \xrightarrow{\mathbb{P}} Q_t$$

Then, due to the stable convergence,

$$\frac{\sqrt{n}}{\sqrt{\widehat{Q}_t^n}}\big(C_t^n - C_t\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Hence, although C_t is random, one can “estimate” it and construct confidence bounds for it.

Estimating the spot volatility σ_t^2. Since $C_t = \int_0^t \sigma_s^2\, ds$ and C_t can be estimated for all t, the quantities σ_t^2 can also be consistently estimated (up to a dt-null set, since anyway one can change σ_t on a dt-null set without changing the process X itself).

In practice, we need some smoothness of $t \mapsto \sigma_t^2$. We have

$$\frac{C_{t+s} - C_t}{s} = \sigma_t^2 + \frac{1}{s} \int_t^{t+s} (\sigma_u^2 - \sigma_t^2)\, du$$

and $\frac{1}{s}\,(\widehat{C}^n_{t+s} - \widehat{C}^n_t) = \frac{1}{s}\,(C_{t+s} - C_t) + \mathsf{O}_P\big(1/\sqrt{ns}\,\big)$. Hence if $t \mapsto \sigma_t^2$ is ρ-Hölder we get

$$n^\theta(\widehat{C}^n_{t+n^{-\theta}} - \widehat{C}^n_t) \;=\; \sigma_t^2 + \mathsf{O}(n^{-\theta\rho}) + \mathsf{O}_P(n^{(\theta-1)/2})$$

and taking $\theta = 1/(1+2\rho)$ gives us

$$n^{1+2\rho}(\widehat{C}^n_{t+1/n^{1+2\rho}} - \widehat{C}^n_t) \;=\; \sigma_t^2 + \frac{1}{n^{\rho(1+2\rho)}}\, \mathsf{O}_P(1)$$

In many (most ?) models we know that $\rho = 1/2$, so the rate is $n^{1/4}$, and the $\mathsf{O}_P(1)$ term above is actually a mixed centered normal variable with an explicit (random) variance.

What can be estimated - 3 ?

The general Itô semimartingale case:

$$X_t = X_0 + \int_0^t b_s\,ds + \int_0^t \sigma_s\,dW_s + \text{sum of jumps}$$

The jumps are driven by a Poisson random measure and characterized (in an appropriate sense) by the family of Lévy measures $(F_t)_{t\in[0,1]}$: each $F_t = F_{t,\omega}(dx)$ is a positive measure on $\mathbb{R}\backslash\{0\}$ such that $\int(1 \bigwedge x^2)F_t(dx) < \infty$, and for any $A \subset \mathbb{R}$ at a positive distance of 0 the process $\sum_{s\leq t} 1_A(\Delta X_s) - \int_0^t F_s(A)\,ds$ is a martingale (here $\Delta X_t = X_t - X_{t-}$).

More concretely, an interpretation of F_t is as follows:

- *If X has finitely many jumps only:* if we know that the stopping time T is a jump time of X, then all F_t are *finite* measures and

$$\frac{F_T(A)}{F_T(\mathbb{R})} = \mathbb{P}(\Delta X_T \mid \mathcal{F}_{T-}) \qquad (\mathcal{F}_{T-} \text{ is the past strictly before } T)$$

- *In general, if T is a jump time of X with size in $D_\varepsilon = \{x : |x| > \varepsilon\}$*

$$\frac{F_T(A\cap D_\varepsilon)}{F_T(D_\varepsilon)} = \mathbb{P}(\Delta X_T \mid \mathcal{F}_{T-})$$

The degree of jump activity. The function $\alpha \mapsto \int_{\mathbb{R}}(|x|^\alpha \bigwedge 1)\, F_t(dx)$ is non-increasing, and

$$\beta_t \;=\; \inf\Big(p \geq 0 : \int_{\mathbb{R}} (|x|^p \bigwedge 1)\, F_t(dx) < \infty\Big)$$

is called the (a priori random) local Blumenthal-Getoor index. Necessarily $\beta_t \in [0,2]$, and the essential supremum $\overline{\beta}_t$ of $s \mapsto \beta_s$ over $[0,t]$ is the (random) global index. We have

$$\begin{array}{lcl} p > \overline{\beta}_t & \Rightarrow & \sum_{S \leq t} |\Delta X_s|^p \;<\; \infty \\ \mathcal{P} < \overline{\beta}_t & \Rightarrow & \sum_{s \leq t} |\Delta X_s|^p \;=\; \infty. \end{array}$$

When X is a Lévy process $\overline{\beta}_t(\omega) = \beta_t(\omega) = \beta$. It also the stability index when X is a stable process.

If the whole path $t \mapsto X_t$ is known:

- All jumps ΔX_t are known, hence $A(p)_t = \sum_{s\leq t} |\Delta X_s|^p$ and $\overline{\beta}_t$ as well.

- The quadratic variation (now equal to $C_t + A(2)_t$ is known, hence C_t as well.

So, in a high-frequency setting we can hope for consistently estimating C_t and $\overline{\beta}_t$, and also σ_t^2 and β_t (if those are smooth enough in t), and also a_t^+ (the “intensity of positive jupms”) if $F_t(x,\infty)) \sim a_t^+/x^\beta$ as $x \downarrow 0$, and the same for negative jumps.

But no consistent inference possible for:

- the drift b_t

- the restriction of the Lévy measures F_t to the complement D_ε of any neighborhood of 0 (that is, for the law of the “big” jumps).

- anything about F_t when F_t is a finite measure for all t.

Test for jumps

This is based on the following (easy) properties, for $p > 2$:

$$V(p)_1^n := \sum_{i=1}^{[nt]} |\Delta_i^n X|^p \xrightarrow{\mathbb{P}} A(p)_1, \qquad n^{p/2-1} V(p)_1^n \xrightarrow{\mathbb{P}} m_p \int_0^1 \sigma_s^p \, ds \text{ on } \Omega^{cont},$$

where Ω^{cont} is the set of all ω for which $t \mapsto X_t(\omega)$ is continuous. When $p = 2$ we have

$$V(2)_1^n = \widehat{C}_1^n \xrightarrow{\mathbb{P}} C_1 + A(2)_1.$$

These convergences enjoy a (not so easy) Central Limit Theorem (CLT) with rate $\sqrt{n}$, so we can define critical (rejection) regions of the form

$$\left\{ \frac{V(p)_1^n}{V(p)_1^{2n}} > 1 + \gamma_n \right\} \qquad \text{for testing the null hypothesis } (\Omega^{Cont})^c$$

$$\left\{ \frac{V(p)_1^n}{V(p)_1^{2n}} < 2^{p/2-1} - \gamma'_n \right\} \qquad \text{for testing the null hypothesis } \Omega^{Cont}$$

Choosing appropriately γ_n or γ'_n, we obtain tests with an asymptotic level α (any prescribed value) and which are asymptotically consistent under the alternative.

Estimating the integrated volatility when jumps occur

When jumps are present, the approximated quadratic variation $\widehat{C}_t^n = \sum_{i=1}^{[nt]} (\Delta_i^n X)^2$ converge to $C_t + A(2)_t$ and not to C_t. For estimating C_t there are (mainly) two methods:

truncation: $$\widehat{C}_t'^n = \sum_{i=1}^{[nt]} (\Delta_i^n X)^2 \, 1_{\{|\Delta_i^n X| \leq u_n\}}$$

multipowers: $$\widehat{C}_t''^n = \alpha_p \sum_{i=1}^{[nt-k+1} \prod_{j=0}^{k-1} |\Delta_{i+j}^n X|^{2/k}$$

Both are consistent, and enjoy a CLT with a centered mixed normal limit (and the efficient variance in the truncation case, a bigger variance for multipowers) provided $\int (|x|^\beta \wedge 1) \mathbb{F}_t(dx)$ is a locally bounded process (this is – slightly – stronger than asking $\overline{\beta}_1 \leq \beta$) for some $\beta < 1$. We can likewise estimate the spot volatility σ_t^2.

Note also that there are tests for the hypotheses

- “The Brownian component is present” ($C_1 > 0$), or “is absent” (i.e., $C_1 = 0$).
- “The jumps have finite activity” ($\overline{\beta}_1 = 0$), or “infinite activity”.

A new approach for estimating integrated volatility and BG index

(with V. Todorov)

Set $\ell_n =$ the integer part of $\log\log n$ and

$$L(u)_n = \sum_{i=1}^{n-\ell_n} \sum_{p=1}^{\ell_n} \frac{1}{p} \prod_{j=0}^{p-1} \big(1-\cos(u\Delta_{i+j}^n X)\big).$$

If $X_t = \sigma W_t + Y_t$, with Y a *symmetric* purely discontinuous Lévy process with a Lévy measure satisfying $\int(|x|\wedge 1)\,F(dx)<\infty$. The variables $\Delta_i^n X$ are i.i.d. and

$$\mathbb{E}\big(\cos(u\Delta_i^n X)\big) = e^{-\sigma^2u^2/2n-\phi(u)/n}, \qquad \text{with } 0\le\phi(u)\le Ku.$$

If $u_n = \sqrt{n}/\log n$, the expectation of the ith summand in $L(u_n)_n$ is thus

$$\sum_{p=1}^{\ell_n} \frac{1}{p}\Big(1-e^{-\sigma^2u_n^2/2n-\phi(u_n)/n}\Big)^p = \frac{\sigma^2u_n^2}{2n} + \mathrm{o}\Big(\frac{u_n}{n} + \big(\frac{u_n^2}{2n}\big)^{\ell_n}\Big)$$

and one can control the variance in a similar way. Then there is no wonder that we get a CLT of the type: $\frac{\sqrt{n}}{u_n^2}\big(L(u_n)_n - \frac{\sigma^2u_n^2}{2}\big)$ converges in law to $\mathcal{N}(0,\sigma^4/2)$.

With a bit more care, the same result holds when $X_t = bt + \sigma W_t + Y_t$ with Y a possibly asymmetric purely discontinuous L´vy process. Now, an Itô semimartingale behaves locally as a Lévy process, with characteristics changing with time. Then (after some calculations), for any Itô semimartingale with the processes b_t and σ_t and $\int (|x| \wedge 1)\, F_t(x)$ locally bounded, we have

$$\frac{\sqrt{n}}{u_n^2}\Big(L(u_n)_n - \frac{u_n^2}{2}\,C_1\Big) \text{ stably converges in law to } \frac{1}{\sqrt{2}}\int_0^1 \sigma_s^2\,dB_s$$

ith B another Brownian motion independent of X. Then a trivial calculation shows us that

$$\sqrt{n}\,\Big(\frac{2L(u_n)_n}{u_n^2} - C_1\Big) \text{ stably converges in law to } \sqrt{2}\int_0^1 \sigma_s^2\,dB_s$$

(so $2L(u_n)_n/u_n^2$ is exactly as good as the truncated estimators introduce before, and both rely on a tuning parameter u_n.

In the symmetric Lévy case $X_t = \sigma W_t + Y_t$ with a Lévy measure satisfying $\int (|x|^\beta \wedge 1)\, F(dx) < \infty$, in the expansion of $L(u_n)_n$ we have a term of size u_n^β/n, which is negligible when $\beta \leq 1$, but not when $\beta > 1$. In the asymmetric case we have another term of typical order $u_n^{\beta\vee 1}/n$ and of course the same is true for Itô semimartingales.

To deal with the asymmetry, we will consider the differences of two successive increments,For the other bias, we need to make a STRONG assumption on the Lévy measures F_t. Namely, the "symmetric tail" $\overline{F}_t = F_t([-x,x]^c)$ satisfies

$$\left| \overline{F}_{t,\omega}(x) - \frac{a_t(\omega)}{x^\beta} \right| \leq \frac{K}{x^{\beta/2}} \qquad \text{for } \; x \in (0,1]$$

(plus some mild but technical requirement) for a constant $\beta \in (0,2)$ and some nonnegative process a_t, smooth enough in t. This in particular implies $\overline{\beta}_1 = \beta$ is non-random on the set $\{\omega : \; A(\omega) = \int_0^1 a_t(\omega) dt > 0\}$ and $\overline{\beta}_t \leq \beta/2$ elsewhere.

This accommodate the case

$$X_t = \int_0^t b_s\, ds + \int_0^t \sigma_s, dW_s + \int_0^t \gamma_{s-}\, dZ_s,$$

with Z a stable or tempered stable process with index β.

Set

$$\mathcal{L}(u)_n = \sum_{i=1}^{n/2-\ell_n} \sum_{p=1}^{\ell_n} \frac{1}{p} \prod_{j=0}^{p-1} \big(1 - \cos(u\Delta^n_{2i+2j-1}X - u\Delta^n_{2i+2j}X)\big).$$

This eliminate the bias due to asymmetry, and the key point about our assumption on F_t is that, in the Lévy case, we have

$$\mathbb{E}\big(\cos(u\Delta^n_i X - u\Delta^n_{i+1}X)\big) = \exp\Big(-\frac{\sigma^2u^2}{n} - \frac{2a\chi_\beta u^\beta}{n} + \mathsf{O}(u^{\beta/2}/n)\Big).$$

So, $\mathcal{L}(u_n)_n$ is an “estimator” for $\frac{u_n^2}{n}\,G(u_n)_n$ where $G(u_n)_n = C_1 + 2\chi_\beta u_n^{\beta-2}A$, Then, to eliminate $u_n^{\beta-2}A$ we use another statistic, where $y > 1$:

$$\mathcal{L}'(u,y)_n = \overline{L}(uy)_n - y^2\overline{L}(y)_n,$$

$$\overline{L}(u)_n = \sum_{i=1}^{n/4-\ell_n} \sum_{p=1}^{\ell_n} \frac{1}{p} \prod_{j=0}^{p-1} \big(1 - \cos(u\Delta^n_{4i+2j-3}X - u\Delta^n_{4i+2j-2}X)\big).$$

We actually have a CLT for $\mathcal{L}(u_n)_n$, and another one for $\mathcal{L}'(u_n, y)$ jointly for all y in a finite set of numbers $y > 1$, under appropriate conditions (depending on $\beta \in (0,2)$) on u_n, and if we put

$$\widehat{C}^n = \frac{1}{u_n^2}\Big(\mathcal{L}(u_n)_n - \frac{\mathcal{L}'(u_n, y)_n^2}{\mathcal{L}'(u_n, y^2) - 2y^2\mathcal{L}'(u_n, y)}\Big),$$

we have

$$\sqrt{n}\big(\widehat{C}^n - C_1\big) \quad \text{converges stably in law to } \; 2\int_0^1 \sigma_s^2 \, dB_s$$

So we loose a factor 2 for the asymptotic variance, but this works for All β.

$\Big($Warning: This HEAVILY depends on the special structure of the Lévy measures F_t; when this assumption fails the result also totally fails...$\Big)$

Estimating the degree of activity of jumps

It is also possible to estimate the degree of activity of jumps (= BG index), but only under the previous restrictive assumption on F_t. Various estimators do the job, but we use the previous method, which leads to “almost” optimal results. We extend the definition of $\mathcal{L}(u)_n$ by putting; for any integer $k \geq 1$:

$$\mathcal{L}(u)_n^k = \sum_{l=1}^{k} C_k^l (-1)^{l+1} \mathcal{L}(u\sqrt{l}_n).$$

Since $\sum_{l=1}^k C_k^l (-1)^{l+1} \, l = 0$ when $k \geq 2$, we see that $\mathcal{L}(u_n)_n^k$ is an “estimator” for $2\chi_\beta \frac{u_n^{\beta-2}}{n} \sum_{l=1}^k C_k^l (-1)^{l+1} \, l^{\beta/2}$, and we do have a CLT for this (again under appropriate conditions on u_n. Then, for $k \geq 2$ and $y > 1$,

$$\widehat{\beta}_n^k = \frac{1}{\log y} \, \log \Big(\frac{\mathcal{L}(u_n y))_n^k}{\mathcal{L}(u_n)_n^k} \Big)$$

and we have

$u_n^{\beta/2} \, \big(\widehat{\beta}_n^k - \beta\big)$ converges stably in law to some centered mixed normal variable

and the variance of the limit can also be estimated ($\Rightarrow$ we can construct confidence intervals).

We need to find u_n “as large as possible” to have a good rate. By taking k large enough, for any $\varepsilon > 0$ we can choose u_n in such a way that

$$n^{\beta/4-\varepsilon}(\widehat{\beta}_n^k - \beta) \quad \text{converges to a centered mixed normal}$$

This has to be compared to the efficient rate which, in the case of a Lévy process with a non-zero Gaussian component is $n^{\beta/4}$ (up to log term).

In a similar way, we can estimate

- The (random) intensity of jumps, that is the variable A.
- The analogue of A for the positive (or negative) jumps only

Finally, when there is no Brownian motion in the picture, the rate improves a lot, and we get for any ε (with a proper choice of k and u_n):

$$n^{1/2-\varepsilon}(\widehat{\beta}_n^k - \beta) \quad \text{converges to a centered mixed normal}$$

When there is noise

.

Now, we suppose that the observations are not $X_{i/n}$, but rather $Y_i^n = X_{T(n,i)} + \varepsilon_i^n$, where ε_i^n is "by definition" the noise. The observed returns are $\Delta_i^n Y = Y_i^n - Y_{i-1}^n$. The assumptions on X are as previous. For the noise, we suppose that the variables $(\varepsilon_i^n : i \geq 1)$ are independent, conditionally on the whole process X, with for all integer $p > 0$:

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

- $\gamma_t^{(0)} \equiv 0$ (the noise is X-conditionally centered)
- $\gamma^{(p)}$ for $p \neq 1$ are smooth enough in t.

An important example: We have i.i.d. variable Z_i^n, independent of X, uniform over $[0, 1]$, and the observation at time i/n is

$$Y_i^n = [X_{i/n} + Z_i^n] \qquad \text{(the integer part)},$$

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

De-noising by pre-averaging

We choose a tuning parameter h_n (an integer) going to ∞ as $n \to \infty$. 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 $g(x) = x^+ \bigwedge (1-x)^+$. 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

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

Then, we can basically re-define the previous statistics $L(u_n)_n$, etc... with $\Delta_i^n X$ substituted with $\widetilde{Y}_i^n$ (with some care: we only take the $\widetilde{Y}_i^n$ for i that are multiples of h_n.

The effect of pre-averaging is to *kill the noise*: as soon as $k_n/\sqrt{n} \to \infty$, $\widetilde{Y}_i^n$ is basically the same as $\widetilde{X}_i^n$ (the pre-averaged value when there is no noise). Then some (tedious) work allows us to see that all previous CLTs are valid, hence the estimators for C_t or β behave as previousle, with one BIG difference: n should be substituted with n/h_n. So the rates we obtain are basically the square-roots of the rates in the no-noise case.

This is of course natural: when there is an additive i.i.d. noise, the optimal rate for estimating C_t for example is no longer $\sqrt{n}$, but $n^{1/4}$.

Extensions

● All what precedes can be done when the sampling is irregular, but relatively close to a kind of “modulated” random walk: that is, the sampling times $T(n,i)$ are no longer i/n, but for each n are a modulated random walk (the variables $T(n,i+1)-T(n,i)$ are, conditionally on X, independent with a law depending “smoothly enough” on X.

● We can consider the case where the Lévy measures have a finite expansion

$$\overline{F}_t(x)=\sum_{m=1}^{M}\frac{a_t^m}{x^{\beta_m}}+\mathrm{o}\Big(\frac{1}{x^{\beta_1/2}}\Big)$$

with $\beta_1>\beta_2>\cdots>\beta_M>\beta_1/2$. It is then possible to estimate the successive indices β_m (with of course decreasing rates as m increases).

Open problems

- Some theoretical problems, such as optimality (efficient rates, efficient asymptotic variances) for the estimations of the various parameters in an Itô semimartingale setting.

- What happens when the BG index β_t is time-varying ?

- What about multi-dimensional processes X (for example, how to de-bias the estimator for the cross integrated volatility, or quadratic co-variation, of the continuous part of two components of X ?).

- How to choose in practice u_n, and h_n in the noisy case (so far, only “mathematical” results are known.)