On Lead-Lag Estimation

Mathieu Rosenbaum
CMAP-École Polytechnique

Joint works with

Marc Hoffmann, Christian Y. Robert and Nakahiro Yoshida

12 January 2011

Outline

Outline

Motivation

Observation from practitioners in finance

- Some assets are leading some other assets.
- This means that a "lagger" asset may partially reproduce the behavior of a "leader" asset.
- This common behavior is unlikely to be instantaneous. It is subject to some time delay called "lead-lag".

A toy model for Lead-Lag

Bachelier model

- For $t \in [0,1]$, and $(B^{(1)}, B^{(2)})$ such that $\langle B^{(1)}, B^{(2)} \rangle_t = \rho t$, set

$$X_t := x_0 + \sigma_1 B_t^{(1)}, \ \widetilde{Y}_t := \widetilde{y}_0 + \sigma_2 B_t^{(2)},$$

- Define $Y_t := \widetilde{Y}_{t-\theta}$, $t \in [\theta, 1]$. Our lead-lag model is given by the bidimensional process (X_t, Y_t).
- We have

$$\begin{cases} X_t &= x_0 + \sigma_1 B_t^{(1)} \\ Y_t &= y_0 + \rho\, \sigma_2 B_{t-\theta}^{(1)} + \sigma_2 (1-\rho^2)^{1/2}\, W_{t-\theta} \end{cases}.$$

Intuitive estimator in the Bachelier model (1)

Estimation idea (1)

- Assume the data arrive at regular and synchronous time stamps in the Bachelier model, *i.e.* we have data

$$(X_0, Y_0), (X_{\Delta_n}, Y_{\Delta_n}), (X_{2\Delta_n}, Y_{2\Delta_n}), \ldots, (X_1, Y_1),$$

  and suppose $\theta = k_0 \Delta_n,\ k_0 \in \mathbb{Z}$.
- Let

$$\mathcal{C}_n(k) := \sum_i \big(X_{i\Delta_n} - X_{(i-1)\Delta_n}\big)\big(Y_{(i+k)\Delta_n} - Y_{(i+k-1)\Delta_n}\big).$$

Intuitive estimator in the Bachelier model (2)

Estimation idea (2)

- Heuristically, we have

$$\mathcal{C}_n(k) \approx \Delta_n^{-1}\mathbb{E}\big[(X_{\cdot} - X_{\cdot-\Delta_n})(Y_{\cdot+k\Delta_n} - Y_{\cdot+(k-1)\Delta_n})\big] + \Delta_n^{1/2}\xi^n.$$

- Moreover,

$$\Delta_n^{-1}\mathbb{E}\big[(X_{\cdot}-X_{\cdot-\Delta_n})(Y_{\cdot+k\Delta_n}-Y_{\cdot+(k-1)\Delta_n})\big] = \begin{cases} 0 & \text{if} \quad k \neq k_0 \\ \rho\,\sigma_1\sigma_2 & \text{if} \quad k = k_0. \end{cases}$$

- Thus we can (asymptotically) detect the value k_0 that defines θ in the very special case $\theta = k_0\Delta_n$ by maximizing in k the contrast sequence

$$k \rightsquigarrow \big|\mathcal{C}_n(k)\big|.$$

Outline

1. Introduction
2. Lead-Lag estimation from non synchronous data
3. Simulations
4. Real Data
5. A random matrices based approach

Covariation estimation

Previous-Tick estimation

- Estimating covariation is an intricate problem as soon as non synchronous data are considered.
- Assume now we observe X at times $(T^{X,i}), i = 1, \dots$ and Y at times $(T^{Y,i}), i = 1, \dots$, with $T^{X,i} \leq T$, $T^{Y,i} \leq T$.
- We build

$$\overline{X}_t = X_{T^{X,i}} \text{ for } t \in [T^{X,i}, T^{X,i+1}),$$

$$\overline{Y}_t = Y_{T^{Y,i}} \text{ for } t \in [T^{Y,i}, T^{Y,i+1}).$$

- For given h, the previous tick covariation estimator is

$$V_h = \sum_{i=1}^{m} \left(\overline{X}_{ih} - \overline{X}_{(i-1)h}\right)\left(\overline{Y}_{ih} - \overline{Y}_{(i-1)h}\right).$$

Drawback of this estimator

Epps effect

- Systematic bias for this estimator.
- Example : Assume that X and Y are two Brownian motions with correlation ρ and that the observation times are arrival times of two independent Poisson processes, then one can show that

$$\mathbb{E}[V_h] \to 0, \text{ as } h \to 0.$$

A convergent estimator under asynchronicity

Hayashi-Yoshida estimator

- Let $I_i^X = (T^{X,i}, T^{X,i+1}]$ and $I_i^Y = (T^{Y,i}, T^{Y,i+1}]$
- The Hayashi-Yoshida estimator is

$$U_n = \sum_{i,j} \Delta X(I_i^X) \Delta Y(I_j^Y) 1_{\{I_i^X \cap I_j^Y \neq \varnothing\}}.$$

- This estimator does not need any selection of h and is convergent.

The lead-lag model

Let $\theta > 0$ (for simplicity, extensions are quite straightforward) and set $\mathbb{F}^{\theta} = (\mathcal{F}_t^{\theta})_{t \geq 0}$, with $\mathcal{F}_t^{\theta} = \mathcal{F}_{t-\theta}$.

Assumptions

- We have
$$X = X^c + A, \quad Y = Y^c + B.$$
- $(X_t^c)_{t \geq 0}$ is a continuous $\mathbb{F}$-local martingale, and $(Y_t^c)_{t \geq 0}$ is a continuous $\mathbb{F}^{\theta}$-local martingale.
- $\exists v_n \to 0,\ v_n^{-1} \max\big\{\sup\{|I_i^X|\}, \sup\{|I_i^Y|\}\big\} \to 0.$
- The $T^{X,i}$ are $\mathbb{F}^{v_n}$-stopping times and the $T^{Y,i}$ are $\mathbb{F}^{\theta + v_n}$-stopping times.

Estimation strategy

Estimator

- We set

$$U_n(\theta) = \sum_{i,j} \Delta X(I_i^X)\Delta Y(I_j^Y)1_{\{I_i^X \cap (I_j^Y)_{-\theta} \neq \varnothing\}},$$

with $(I_j^Y)_{-\theta} = (T^{Y,j} - \theta, T^{Y,j+1} - \theta]$.

- Eventually, $\widehat{\theta}_n$ is defined as a solution of

$$\big|U^n(\widehat{\theta}_n)\big| = \max_{\theta \in \mathcal{G}^n} \big|U^n(\theta)\big|,$$

where $\mathcal{G}^n$ is a sufficiently fine grid.

Result

Theorem

As $n \to \infty$,

$$v_n^{-1}\left(\widehat{\theta}_n - \theta\right) \to 0,$$

in probability, on the event $\left\{\langle X^c, \widetilde{Y}^c\rangle_T \neq 0\right\}$.

Outline

1. Introduction
2. Lead-Lag estimation from non synchronous data
3. Simulations
4. Real Data
5. A random matrices based approach

Synchronous case

Setup

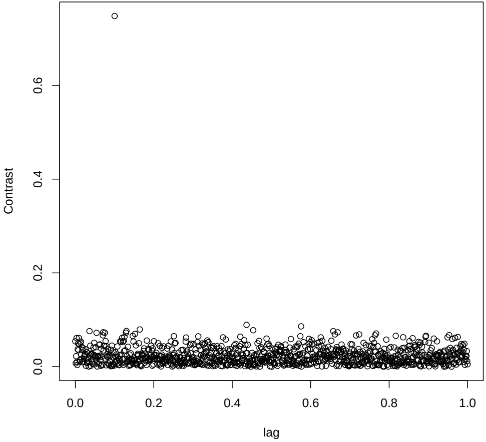
- We consider 300 simulations of the Bachelier model with synchronous equispaced data with period Δ_n.
- $t \in [0,1],\ \theta = 0.1,\ x_0 = \widetilde{y}_0 = 0,\ \sigma_1 = \sigma_2 = 1.$
- The mesh size of the grid h_n satisfies $h_n = \Delta_n$.
- We consider the following variations :
  - Mesh size : $h_n \in \{10^{-3}(FG), 3.\,10^{-3}(MG), 6.\,10^{-3}(CG)\}$.
  - Correlation value : $\rho \in \{0.25, 0.5, 0.75\}$.

Results in the synchronous case

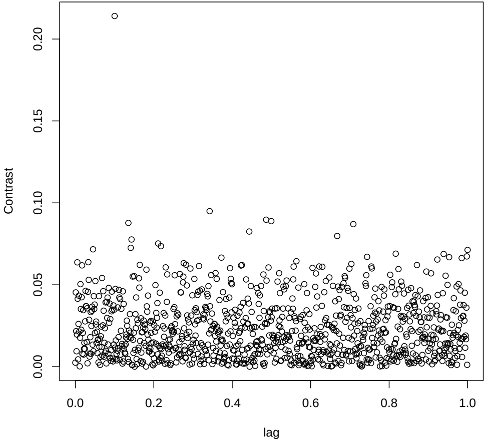
$\widehat{\theta}_n$	0.096	0.099	0.1	0.102	Other
FG, $\rho = 0.75$	0	0	300	0	0
MG, $\rho = 0.75$	0	300	0	0	0
CG, $\rho = 0.75$	1	0	0	299	0
FG, $\rho = 0.50$	0	0	300	0	0
MG, $\rho = 0.50$	0	299	0	1	0
CG, $\rho = 0.50$	13	0	0	280	7
FG, $\rho = 0.25$	0	0	300	0	0
MG, $\rho = 0.25$	0	152	0	11	137
CG, $\rho = 0.25$	10	0	0	66	124

Table 1 : *Estimation of* $\theta = 0.1$ *on 300 simulated samples for* $\rho \in \{0.25, 0.5, 0.75\}$.

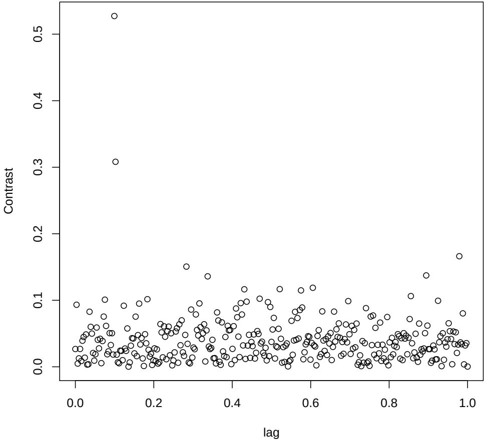
One sample path, FG, $\rho = 0.75$



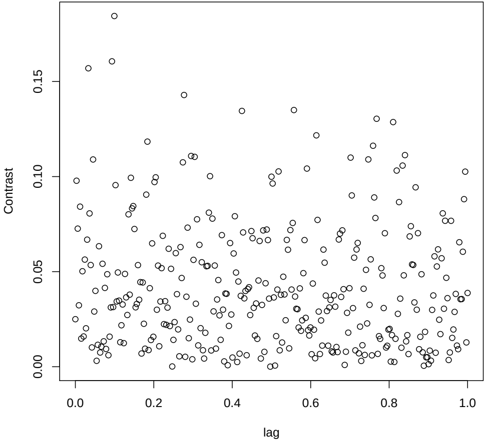
One sample path, FG, $\rho = 0.25$



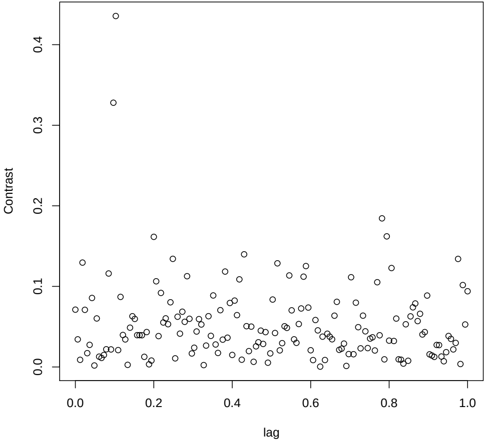
One sample path, MG, $\rho = 0.75$



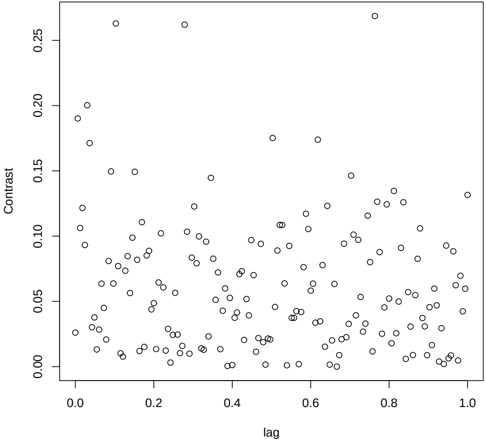
One sample path, MG, $\rho = 0.25$



One sample path, CG, $\rho = 0.75$



One sample path, CG, $\rho = 0.25$



Non synchronous case

Setup

- We randomly pick 300 sampling times for X over $[0, 1]$, uniformly over a grid of mesh size 10^{-3}.
- We randomly pick 300 sampling times for Y likewise, and independently of the sampling for X.
- Fine grid case, with $\theta = 0.1$ and $\rho = 0.75$.

Results for the non synchronous case

$\widehat{\theta}$	0.099	0.1	0.101	0.102	0.103	0.104	0.105
FG, $\rho = 0.75$	16	106	107	46	19	4	2

Table 2 : *Estimation of $\theta = 0.1$ on 300 simulated samples for $\rho = 0.75$ and non-synchronous data.*

Outline

1. Introduction
2. Lead-Lag estimation from non synchronous data
3. Simulations
4. Real Data
5. A random matrices based approach

The data

Dataset

We study here the lead-lag relationship between the two following assets :

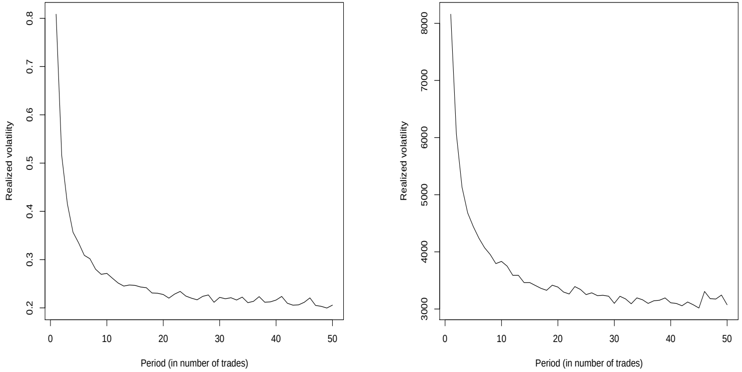
- The future contract on the DAX index, with maturity December 2010,
- The Euro-Bund future contract, with maturity December 2010.

Dealing with microstructure noise

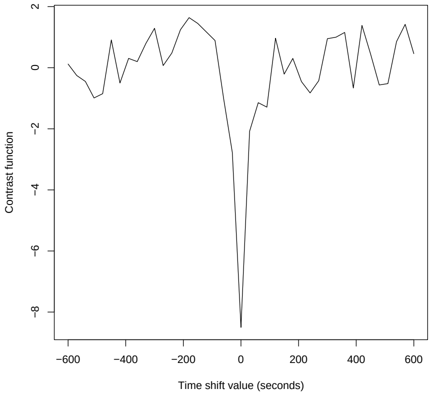
Methodology

- We want to use high frequency data.
- First approach : use of the Uncertainty Zones Model.
- Here we just use signature plots in trading times. This enables to take advantage of non synchronous data.
- We keep one trade out of twenty.
- We then compute the function U^n over these trades.

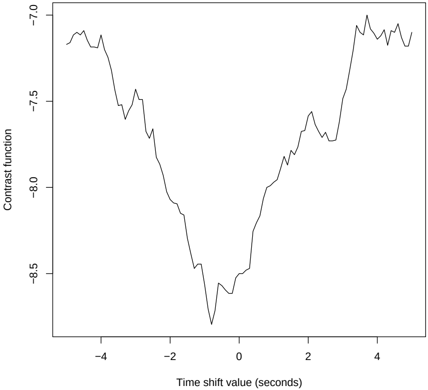
Signature plots, October 13, for Bund (left) and FDAX (right).



Function U^n, October 13, between -10 and 10 minutes, mesh=30 seconds



Function U^n, October 13, between -5 and 5 seconds, mesh=0.1 second



Bund and DAX, lead-lag estimation

Jour	Vol.(Bund)	Vol.(FDAX)	LL.	J.	Vol.B.	Vol.F.	LL
1 Oct.	2847	4215	-0.2	18 Oct.	1727	2326	-2.1
5 Oct.	2213	3302	-1.1	19 Oct.	2527	3162	-1.6
6 Oct.	2244	2678	-0.1	20 Oct.	2328	2554	-0.5
7 Oct.	1897	3121	-0.5	21 Oct.	2263	3128	-0.1
8 Oct.	2545	2852	-0.6	22 Oct.	1894	1784	-1.2
11 Oct.	1050	1497	-1.4	25 Oct.	1501	2065	-0.4
12 Oct.	2265	3018	-0.8	26 Oct.	2049	2462	-0.1
13 Oct.	2018	3037	-0.8	27 Oct.	2606	2864	-0.6
14 Oct.	2057	2625	0.0	28 Oct.	1980	2632	-1.3
15 Oct.	2571	3269	-0.7	29 Oct.	2262	2346	-1.6

Outline

Model (1)

Dynamics

We consider two processes $(X_t)_{t\in[0,1]}$ (leader) and $(Y_t)_{t\in[0,1]}$ (lagger) such that

$$X_t - X_0 = \int_0^t K_{s+\theta} \mathrm{d}W_{s+\theta},$$

$$Y_t - Y_0 = \rho \int_0^{t\wedge\theta} K_s \mathrm{d}\tilde{W}_s + \rho \int_\theta^{t\vee\theta} K_s \mathrm{d}W_s + \int_0^t L_s \mathrm{d}W'_s.$$

- The interval $[\theta, 1]$ is the set of time where the lead-lag relation is in force.
- For $s \in [\theta, 1]$,

$$\mathrm{d}Y_s = \rho \mathrm{d}X_{s-\theta} + L_s \mathrm{d}W'_s.$$

Model (2)

Observations

- We consider $m+1$ equidistant data for each process : $(X_{i/m}, Y_{i/m})$, for $i = 0, \dots, m$.
- $m = p\lfloor p^a \rfloor$ where p is a positive integer and $a > 0$.
- Later, p will be the order of magnitude of the number of "days" the processes will be observed and $m+1$ the number of data per day. This parameter will drive the asymptotic.

Increments

We consider increments of the processes on grids with mesh $1/p$.

Notation

For $i = 1, \dots, p$, and $l = 0, \dots, \lfloor p^a \rfloor$

- $\Delta^{(l,p)} X_i = X_{i/p+l/m} - X_{(i-1)/p+l/m}$.
- $\Delta^{(0,p)} X_i$ and $\Delta^{(l,p)} Y_i$ are centered Gaussian with variance

$$v_{i,0}^X = \int_{(i-1)/p}^{i/p} K_{s+\theta}^2 \mathrm{d}s, \quad v_{i,l}^Y = \int_{(i-1)/p+l/m}^{i/p+l/m} (\rho^2 K_s^2 + L_s^2)\mathrm{d}s.$$

- Random vector of interest :

$$Z^{(l,p)} = p^{1/2}\big(\Delta^{(0,p)} X_1, \dots, \Delta^{(0,p)} X_p, \Delta^{(l,p)} Y_1, \dots, \Delta^{(l,p)} Y_{p-1}\big)^\top.$$

Theoretical covariance

Let $\lfloor \theta \rfloor_p = \lfloor p\theta \rfloor / p$. $Z^{(l,p)}$ is a Gaussian vector of size $2p-1$ with 5-diagonal covariance matrix $\Sigma_{(l,p)}$. For $l = 0, \ldots, \lfloor m(\theta - \lfloor \theta \rfloor_p) \rfloor$:

$$\begin{cases} 1 \leq i \leq p,\ 1 \leq j \leq p,\ i = j & (\Sigma_{(l,p)})_{i,j} = pv^X_{i,0} \\ p+1 \leq i \leq 2p-1,\ p+1 \leq j \leq 2p-1,\ i = j & (\Sigma_{(l,p)})_{i,j} = pv^Y_{j-p,l} \\ 1 \leq i \leq p,\ p+1 \leq j \leq 2p-1,\ j-p = i + p\lfloor \theta \rfloor_p & (\Sigma_{(l,p)})_{i,j} = pv^{XY}_{i,l,1} \\ 1 \leq i \leq p,\ p+1 \leq j \leq 2p,\ j-p = i + p\lfloor \theta \rfloor_p + 1 & (\Sigma_{(l,p)})_{i,j} = pv^{XY}_{i,l,2} \\ p+1 \leq i \leq 2p-1,\ 1 \leq j \leq p,\ i-p = j + p\lfloor \theta \rfloor_p & (\Sigma_{(l,p)})_{i,j} = pv^{XY}_{j,l,1} \\ p+1 \leq i \leq 2p-1,\ 1 \leq j \leq p,\ i-p = j + p\lfloor \theta \rfloor_p + 1 & (\Sigma_{(l,p)})_{i,j} = pv^{XY}_{j,l,2} \end{cases}$$

with

$$v^{XY}_{i,l,1} = \rho \int_{(i-1)/p}^{i/p-(\theta-\lfloor \theta \rfloor_p)+l/m} K^2_{s+\theta} ds \text{ and } v^{XY}_{i,l,2} = \rho \int_{i/p-(\theta-\lfloor \theta \rfloor_p)+l/m}^{i/p} K^2_{s+\theta} ds.$$

- The parameter θ appears in the location of the diagonals.
- Analogous result for $l = \lfloor m(\theta - \lfloor \theta \rfloor_p) \rfloor + 1, \ldots, \lfloor p^a \rfloor$.

Result

Theorem

Using random matrices theory results, we can build another estimator of the lead-lag parameter and provide its asymptotic theory.