# On Lead-Lag Estimation

Mathieu Rosenbaum
CMAP-École Polytechnique Paris

Joint works with

Marc Hoffmann, Christian Y. Robert and Nakahiro Yoshida

17 March 2009

# Outline

1. Introduction-Model
2. Idea of the estimation procedure
3. Results based on Random Matrix Theory
4. Construction of the estimator

# Outline

1. Introduction-Model
2. Idea of the estimation procedure
3. Results based on Random Matrix Theory
4. Construction of the estimator

# Motivation

## Remarks from practitioners

- Some assets are leading some other assets.
- This means that a “lager” 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 the lead-lag delay.

## Objective and methods

- The goal of this work is to model this lead-lag relation and to give an estimator of the lead lag parameter.
- Our approach mainly uses Random Matrix Theory.

# Observations

We considered two assets with log prices $(X_t, Y_t)$.

- $n$ days of data are available.
- we have $m$ equidistant prices for each assets everyday.

So, we observe for $i = 0, \ldots, m-1$ and $d = 1, \ldots, n$

$$\big(X_{i/m+(d-1)}, Y_{i/m+(d-1)}\big).$$

We assume

- $m = p\lfloor p^a \rfloor$ with $p$ a positive integer and $0 < a < 1$.
- $\gamma_p = n/(2p) \underset{p \to +\infty}{\to} \gamma > 0$.

# Dynamics

Let

$$\mathcal{U}(t,\theta) = \big\{ s \leq t, s \in \{[d+\theta, d+1], d \in \mathbb{N}\} \big\}$$
$$\mathcal{V}(t,\theta) = \big\{ s \leq t, s \in \{[d, d+\theta], d \in \mathbb{N}\} \big\}.$$

We assume the log-prices of our two assets $X_t$ (leader) and $Y_t$ (lager) satisfy for some $\theta_p$

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

$$Y_t - Y_0 = \rho \int_{\mathcal{U}(t,\theta_p)} K_s \mathrm{d}W_s + \rho \int_{\mathcal{V}(t,\theta_p)} K_s \mathrm{d}\tilde{W}_s + \int_0^t L_s \mathrm{d}W'_s.$$

## Lead-lag relation

In particular, for $d \in \mathbb{N}$ and $\theta_p \leq t \leq 1$,

$$Y_{d+t} - Y_{d+\theta_p} = \rho(X_{d+t-\theta_p} - X_d) + \int_{d+\theta_p}^{d+t+\theta_p} L_s \mathrm{d}W'_s.$$

# Assumptions

- The functions $K$ and $L$ are deterministic 1-periodic.
- There exists $\theta \in [0,1]$ such that

$$\frac{\lfloor p\theta \rfloor}{p} \leq \theta_p \leq \frac{\lfloor p\theta \rfloor + 1}{p}.$$

- $\theta_p - \mathcal{G}_p(\theta_p) = o(1/p^2)$, where $\mathcal{G}_p(\theta_p)$ is the nearest value from $\theta_p$ on the grid with mesh $1/p$.

This last technical assumption will enable us to prove the result assuming $\theta_p = \lfloor p\theta \rfloor / p$.

# Outline

1. Introduction-Model
2. Idea of the estimation procedure
3. Results based on Random Matrix Theory
4. Construction of the estimator

# Increments

For $i = 1, \dots, p$ and $d = 1, \dots, n$, we define

$$\begin{aligned} \Delta^{(p)} X_i^{(d)} &= X_{i/p+(d-1)} - X_{(i-1)/p+(d-1)} \\ \Delta^{(p)} Y_i^{(d)} &= Y_{i/p+(d-1)} - Y_{(i-1)/p+(d-1)}. \end{aligned}$$

These random variables are independent centered Gaussian with respective variance $v_i^X$ and $v_i^Y$ satisfying

$$v_i^X = \int_{(i-1)/p}^{i/p} H_s^2 \mathrm{d}s, \quad v_i^Y = \int_{(i-1)/p}^{i/p} (\rho^2 K_s^2 + L_s^2) \mathrm{d}s,$$

with $H_s = K_{s+\theta_p}$.

# Covariance Matrix

Now, for $d = 1, \dots, n$, we set $Z_d^{(p)} = (X_d^{(p)}, Y_d^{(p)})$ with

$$X_d^{(p)\top} = (p^{1/2}\Delta^{(p)} X_i^{(d)})_{i=1,\dots,p}, \quad Y_d^{(p)\top} = (p^{1/2}\Delta^{(p)} Y_i^{(d)})_{i=1,\dots,p}.$$

Let

$$v_i^{XY} = \rho \int_{i/p}^{(i-1)/p} H_s^2 \mathrm{d}s.$$

The vector $Z_d^{(p)\top}$ is a Gaussian vector with covariance matrix $\Sigma_p \in \mathbb{R}^{2p} \times \mathbb{R}^{2p}$ such that

$$\begin{cases} 1 \leq i \leq p,\, 1 \leq j \leq p,\, i = j & (\Sigma_p)_{i,j} = p v_i^X \\ p+1 \leq i \leq 2p,\, p+1 \leq j \leq 2p,\, i = j & (\Sigma_p)_{i,j} = p v_{i-p}^Y \\ 1 \leq i \leq p,\, p+1 \leq j \leq 2p,\, j-p = i + p\theta_p & (\Sigma_p)_{i,j} = p v_i^{XY} \\ p+1 \leq i \leq 2p,\, 1 \leq j \leq p,\, i-p = j + p\theta_p & (\Sigma_p)_{i,j} = p v_j^{XY} \end{cases}$$

with all the other terms equal to zero.

# Population spectral distribution

We call $H_p$ the population spectral distribution, that is the distribution that puts mass $1/(2p)$ at each of the eigenvalues of the covariance matrix $\Sigma_p$. We have the following proposition.

## Proposition

The population spectral distribution $H_p$ converges weakly to a distribution denoted by $H_\infty$. Let

$$\mu_{k,p} = \frac{1}{2p} Tr(\Sigma_p^k) = \int \lambda^k \mathrm{d} H_p(\lambda).$$

For any positive integer $k$, we moreover have

$$\mu_{k,p} \underset{p \to +\infty}{\longrightarrow} \mu_k = \int \lambda^k \mathrm{d} H_\infty(\lambda).$$

## Moments of order 1 and 2

The proofs of these results use Carleman condition together with the method of moments. In particular we have

$$\mu_{1,p} = \frac{1}{2p} Tr(\Sigma_p) = \int_0^1 (H_s^2 + \rho^2 K_s^2 + L_s^2) \mathrm{d}s$$

and

$$\mu_{2,p} = \frac{1}{2p} Tr(\Sigma_p^2)$$

$$= \int_0^1 \left(H_s^4 + (\rho^2 K_s^2 + L_s^2)^2\right) \mathrm{d}s + 2\rho^2 \int_0^{(1-\theta_p)} H_s^4 \mathrm{d}s + o(1/p).$$

Idea : using some kind of empirical counterpart to this trace and deriving $\theta_p$ from it.

# Outline

1 Introduction-Model

2 Idea of the estimation procedure

3 Results based on Random Matrix Theory

4 Construction of the estimator

# Empirical counterpart

We now consider the empirical counterpart of $\Sigma_p$ given by $S_p = Z^\top Z/n$, where $Z$ is the matrix in $\mathbb{R}^n \times \mathbb{R}^{2p}$ defined by $Z^\top = (Z_1^{(p)\top}, \dots, Z_n^{(p)\top})$. We call $F_p$ the (random) population spectral distribution of $S_p$. For a distribution $G$, we denote by $m_G$ its Stieltjes transform, that is

$$m_G(z) = \int \frac{\mathrm{d}G}{x - z}, \text{ for } z \in \mathbb{C}^+.$$

and set

$$v_G(z) = -(1 - 2\gamma)\frac{1}{z} + \gamma m_G(z).$$

# Marcenko-Pastur

The following proposition is derived from the Marcenko-Pastur result, see Silverstein (95), Yin and Krishnaiah (83).

## Marcenko-Pastur

Almost surely, $F_p$ converges weakly to a (non random) distribution $F_\infty$. This distribution is characterized by the Marcenko-Pastur type equation

$$-\frac{1}{v_{F_\infty}(z)} = z - \gamma \int \frac{\lambda \mathrm{d} H_\infty(\lambda)}{1 + \lambda v_{F_\infty}(z)}.$$

## Moment of order 2

For any positive integer $k$, we get

$$\frac{1}{2p} Tr(S_p^k) = \int \lambda^k \mathrm{d}F_p(\lambda) \xrightarrow[p \to +\infty]{L^2} m_k,$$

where $m_k = \int \lambda^k \mathrm{d}F_\infty(\lambda)$. Moreover, we obtain

$$\frac{1}{2p} Tr(S_p^2) = \gamma_p \mu_{1,p}^2 + \mu_{2,p} + \mathcal{O}(1/p).$$

## A moment relationship

Recall that

$$\mu_{1,p} = \int_0^1 (H_s^2 + \rho^2 K_s^2 + L_s^2)\mathrm{d}s$$

$$\mu_{2,p} = \int_0^1 \left(H_s^4 + (\rho^2 K_s^2 + L_s^2)^2\right)\mathrm{d}s + 2\rho^2 \int_0^{(1-\theta_p)} H_s^4 \mathrm{d}s + o(1/p).$$

Thus, the relation

$$\frac{1}{2p} Tr(S_p^2) = \gamma_p \mu_{1,p}^2 + \mu_{2,p} + \mathcal{O}(1/p)$$

gives us a moment relationship for estimating $\theta_p$.

# Outline

1 Introduction-Model

2 Idea of the estimation procedure

3 Results based on Random Matrix Theory

4 Construction of the estimator

# Estimation of functionals

## Variations

To use the relationship, we need estimates for the quantities

$$\int_0^t H_s^2 \mathrm{d}s, \int_0^t H_s^4 \mathrm{d}s, \int_0^t (\rho^2 K_s^2 + L_s^2) \mathrm{d}s, \int_0^t (\rho^2 K_s^2 + L_s^2)^2 \mathrm{d}s.$$

Their estimators $\hat{V}_t^X$, $\hat{Q}_t^X$, $\hat{V}_t^Y$ and $\hat{Q}_t^Y$ are estimated using the grid with mesh $1/m$, thanks to averages of daily quadratic variations and quarticity. The rate of convergence obtained is faster than $p$.

# Estimation of $\rho$

## Thresholded estimator

We assume there exists a known constant $c > \theta$. Let for $k \leq mc$,

$$\Gamma_k^{(d)} = \sum_{1 \leq j \leq m(1-c)} \Delta^{(m)} X_j^{(d)} \Delta^{(m)} Y_{j+k}^{(d)}.$$

We define

$$\hat{\rho} = \frac{1}{n \hat{V}_{1-c}^X} \sum_{d=1}^{n} \sum_{0 \leq k \leq mc} \Gamma_k^{(d)} 1_{\{|\Gamma_k^{(d)}| \geq s(m,a)\}},$$

# Definition of the estimator

## Estimator of $\theta_p$

We set

$$A_p = \frac{1}{p} Tr(S_p^2) - \gamma_p(\hat{V}_1^X + \hat{V}_1^Y)^2 - (\hat{Q}_1^X + \hat{Q}_1^Y).$$

We define $\hat{\theta}_p$ by

$$\hat{\theta}_p = \min\big\{\theta_p \in \{\frac{k}{m},\ k = 1, \dots, m\}, 2\hat{\rho}^2 \hat{Q}^X_{1-\theta_p} \leq A_p\big\}.$$

# Limit theorem for lead-lag estimation

The following limit theorem for $\hat{\theta}_p$ is obtained using a result from Bai and Silverstein (04).

**Theorem**

The sequence

$$p(\hat{\theta}_p - \theta_p)$$

weakly converges to a Gaussian random variable.

# Discussion

## Comments

- In this result, $1/p$ has to be understood as the time between two data and not as the inverse of the number of data.
- The accuracy of our estimator is comparable to those of the estimator of Hoffmann, R., Yoshida. This other estimator is based on the non synchronous covariation estimator of Hayashi and Yoshida (03) and obtained in a high frequency context (one day of data).
- Indeed, the way the number of data increases the information on $\theta$ is very intricate. Because of the uncertainty on the correlation parameter, the accuracy of the estimator of H., R., Y. is probably not improved by adding days of data.
- The random matrix based estimator enables (probably) to treat the case of a gaussian microstructure noise.