Detecting infinitesimal lead-lag effects from noisy high-frequency data

Yuta Koike

The Institute of Statistical Mathematics

and

CREST JST

S.A.P.S. X, Le Mans, March 17, 2015

Outline

- Introduction
- Model
- Formulation of the problem
- Main results
- Simulation study
- Conclusions

Introduction

- Lead-lag effect
  - One process (“leader”) is correlated with another process (“lagger”) at later times
- The investigation of such a relationship has a long history in economics
- Classically, it has been examined at moderate frequencies (day, week, month, quarter, . . .) using the statistics for discrete-time (stationary) processes
  - Ex. spectral analysis (cf. Granger & Hatanaka, 1964), distributed lags (cf. Griliches, 1967), cross-autocorrelations (cf. Campbell *et al.*, 1997), . . .

Introduction

- Recently, lead-lag effects at (ultra) high-frequencies have begun to attract notice (e.g. Huth & Abergel, 2014)
- For high-frequency data, discrete-time process modeling tends to be poor; a discretely observed continuous-time process is often more appropriate
- However, there are not many theoretical results on the statistical inference for lead-lag effects in such a setting
- The aim of this talk is to contribute to this area

Introduction

- There are a few approaches to express lead-lag effects
  - <u>Hoffmann, Rosenbaum & Yoshida (2013)</u>

    Model: continuous semimartingale

    Estimation: Hayashi-Yoshida estimator
  - <u>Robert & Rosenbaum (2010)</u>

    Model: continuous Gaussian martingale

    Estimation: random matrix theory
  - <u>Bacry, Delattre, Hoffmann & Muzy (2013)</u>

    Model: Hawkes process

    Estimation: parameter estimation

Introduction

- This talk focuses on the Hoffmann-Rosenbaum-Yoshida model and investigates
  - how to deal with observation noise
  - how to detect “small” lags
- In particular, we will provide a simple but effective hypothesis testing procedure to detect a small lead-lag effect
- We only consider a simple model; an extension to the general case would be possible (in progress)

Model

- $(X_t^1 \; X_t^2)_{t\in[0\;1]}$: bivariate Brownian motion with a lead-lag effect

$$X_t^1 = \quad {}_1 B_t^1 \qquad X_t^2 = \quad {}_2 B_{t-}^2$$

  - $B_t = (B_t^1 \; B_t^2)$, $t \in \mathbb{R}$: two-sided bivariate standard Brownian motion with correlation $\rho \neq 0$ such that $B_0 = 0$
  - $\quad {}_1 \quad {}_2 \quad 0; \quad \in \mathbb{R}$ is the lag parameter

- We observe X at equidistant times with noise: $Y_0^p = 0$ and

$$Y_i^p = X_{t_i}^p + \epsilon_i^p \qquad t_i = i \; n \qquad (i = 1 \qquad n) \tag{1}$$

  - $\epsilon_i^p \overset{i\,i\,d}{\sim} N(0 \; \Upsilon_p)$ and ϵ^1 and ϵ^2 are mutually independent

Model

- We are interested in the inference for the parameter
- We restrict our attention to the situation where the lag is nearly zero

  $\Longrightarrow$ We consider the local asymptotics such that $:= \;\; _n = c\eta_n$ for some $c \in [-\delta_c \;\; \delta_c]$ and $\eta_n \to 0$ as $n \to \infty$
- Empirically, the sizes of lags are usually comparable with the sampling frequency, so such a setting is meaningful
- We assume $\eta_n = o(n^{-\frac{1}{2}})$; we show that this setting allows us to construct a simple, feasible and rate-optimal test for the absence of a lead-lag effect

Formulation of the problem

- We consider the following hypothesis testing problem:

$$H_0 : c = 0 \qquad \text{vs.} \qquad H_1 : c \neq 0 \tag{2}$$

- To discuss the rate optimality problem, we employ the minimax approach of Ingster (1993)
- Namely, we seek the fastest rate $r_n \to 0$ such that the hypothesis testing problem

$$H_0 : c = 0 \qquad \text{vs.} \qquad H_1(r_n) : c \in \mathcal{C}(r_n) \tag{3}$$

permits a uniformly consistent test, where

$$\mathcal{C}(r_n) = \{c : r_n \leq |c| \leq \delta_c\}$$

Formulation of the problem

- In terms of , (3) can be rewritten as

$$H_0 : \quad = 0 \qquad \text{vs.} \qquad H_1(r_n) : r_n\eta_n \leq | \quad | \leq \delta_c\eta_n$$

$\Longrightarrow$ Therefore, our aim corresponds to seeking the fastest rate r_n such that the lag $\quad = r_n\eta_n$ is distinguishable from 0

- The formal formulation of the problem is given in the next slides:

□ Notation

- $\mathcal{E}_n = (\mathcal{X}_n \ \ \mathcal{A}_n \ \ (P_{n\ c})_{c\in[-\delta_c\ \delta_c]})$: our statistical experiments
- Ψ_n: the set of all tests at the stage n, i.e.

$$\in \Psi_n \quad \Leftrightarrow \qquad : \mathcal{X}_n \to \{0 \ \ 1\} \text{ is } \mathcal{A}_n\text{-measurable}$$

  - ▽ $\quad = 0 \Rightarrow H_0$ is accepted
  - ▽ $\quad = 1 \Rightarrow H_0$ is rejected

- $\alpha_n(\ \ ) = P_{n\ 0}(\quad = 1)$: type I error probability for (3)
- $\beta_n(\quad r_n) = \sup_{c\in\mathcal{C}(r_n)} P_{n\ c}(\quad = 0)$: maximal type II error probability for (3)
- $\gamma_n(r_n) = \inf_{\ \in\Psi_n} \{\alpha_n(\ \ ) + \beta_n(\quad r_n)\}$: minimax total error probability for (3)

Definition 1 (Ingster, 1993; Spokoiny, 1996)

A sequence r_n^* is called the **minimax rate of testing** if $r_n^* \to 0$ and

(i) For any sequence r_n such that $r_n = o(r_n^*)$ we have $\gamma_n(r_n) \to 1$,

(ii) For any $\alpha\ \beta \quad 0$, there is a constant $K \quad 0$ and a sequence $\ _n \in \Psi_n$ of tests such that

$$\limsup_{n\to\infty} \alpha_n(\ _n) \leq \alpha \qquad \limsup_{n\to\infty} \beta_n(\ _n\ K r_n^*) \leq \beta$$

Warm-up: an idealized case

- As a warm-up, we consider the idealized situation such that $_1 = \ _2 = 1$ and ρ is known
- We start with the case that the noise is absent

Proposition 1

If $\Upsilon_1 = \Upsilon_2 = 0$ and $|\rho| \quad 1$, the minimax rate of testing for (2) is $r_n^* = n^{-\frac{3}{2}}\eta_n^{-1}$, provided that $r_n^* \to 0$

- If $\rho = 1$ (resp. $\rho = -1$), H_0 is equivalent to saying $X_t^1 = X_t^2$ (resp. $X_t^1 = -X_t^2$) for all t, so any lag is detectable

- A rate-optimal test is constructed based on the fact that lead-lag effects cause the Epps effects

- More formally, if $\neq 0$ and it does not depend on n, the realized covariance $U_n = \sum_{i=1}^{n}(X^1_{t_i} - X^1_{t_{i-1}})(X^2_{t_i} - X^2_{t_{i-1}})$ tends to 0 as $n \to \infty$

- On the other hand, $\sqrt{n}(U_n - \rho) \xrightarrow{d} N(0 \;\; 1 + \rho^2)$ if $= 0$

- This suggests the test rejecting H_0 if $|T_n| \quad z_{1-\alpha \;\; 2}$, where

$$T_n = \sqrt{n}\frac{U_n - \rho}{\sqrt{1 + \rho^2}}$$

and $z_{1-\alpha \;\; 2}$ is the $(1 - \alpha \;\; 2)$-quantile of $N(0 \;\; 1)$

- The following proposition due to Hoffmann *et al.* (2013) ensures that the above test indeed satisfies condition (ii) of Definition 1:

Proposition 2 (Hoffmann *et al.*, 2013, Proposition 1)

Assume $\Upsilon_1 = \Upsilon_2 = 0$. Then we have

$$U_n = \rho\phi(n\ \ ) + n^{-\frac{1}{2}}\sqrt{1+\rho^2\phi(n\ \ )}\ _n$$

under $P_{n\ c}$ for all $n\ \ c$, where $\phi(t) = (1-|t|)1_{\{|t|\leq 1\}}$ and $_n$ is a random variable with zero mean and unit variance and converges in law to $N(0\ \ 1)$ as $n \to \infty$ (under $P_{n\ c}$ for all $n\ \ c$).

Minimax optimality in the noisy case

- We turn to the noisy case

Theorem 1

The minimax rate of testing for (2) is $r_n^* = n^{-\frac{3}{4}}\eta_n^{-1}$, provided that $r_n^* \to 0$

- This result is "canonical" in the sense that the smallest detectable lag $r_n^*\eta_n = n^{-\frac{3}{4}} (= (\sqrt{n})^{-\frac{3}{2}})$ coincides with the one in the non-noisy case with the sample size $\sqrt{n}$ (cf. Gloter & Jacod, 2001)

Construction of a rate-optimal test

- A natural idea is to consider a pre-averaged version of T_n (cf. Podolskij & Vetter, 2009; Vetter & Dette, 2012)

- Namely, we replace U_n with $\overline{U}_n = \frac{1}{k_n} \sum_{i=0}^{n-k_n+1} \overline{Y}_i^1 \overline{Y}_i^2$, where $= 1 \;\; 12$ and

$$\overline{Y}_i = \frac{1}{k_n} \left(\sum_{p=0}^{k_n \;\, 2-1} Y_{i+p+k_n \;\, 2} - \sum_{p=0}^{k_n \;\, 2-1} Y_{i+p} \right)$$

with k_n being a positive even integer s.t. $k_n = \theta \sqrt{n} + o(n^{1 \;\; 4})$ for some $\theta \quad 0$

Construction of a rate-optimal test

- According to Theorem 2 of Christensen, Kinnebrock & Podolskij (2010), $n^{1\ 4}(\overline{U}_n - \rho) \xrightarrow{d} N(0\ \Gamma)$ for some constant $\Gamma \quad 0$ if $\quad = 0$

- Indeed, $\overline{U}_n$ is “too stable” for our purpose:

Proposition 3

If $\quad = o(n^{-5\ 8})$, $n^{1\ 4}(\overline{U}_n - \rho) \xrightarrow{d} N(0\ \Gamma)$ as $n \to \infty$

- The proposition implies that the tests based on the statistic $n^{1\ 4}(\overline{U}_n - \rho) \quad \sqrt{\Gamma}$ cannot detect lags smaller than $n^{-5\ 8}$

Construction of a rate-optimal test

- We suppose $\in \{k\ n : k \in \mathbb{Z}\}$ for simplicity
- Fourier coefficients of $\mathrm{d}X$ (cf. Malliavin & Mancino, 2009):

$$c_f(\mathrm{d}X) = \sum_{i=1}^{n} \exp(-2\pi f \sqrt{-1} t_i)(X_{t_i} - X_{t_{i-1}})$$

- Since $E[(X^1_{t_i} - X^1_{t_{i-1}})(X^2_{t_i} - X^2_{t_{i-1}})] = \rho\ n$ if $t_j - t_i =$ and it vanishes otherwise, we have

$$E[c_f(\mathrm{d}X^1)c_{-f}(\mathrm{d}X^2)] = \exp(2\pi f \sqrt{-1}\ )\rho$$

ignoring the end effects

- This suggests that (a functional of)   would be estimated by smoothing $c_f(\mathrm{d}X^1)c_{-f}(\mathrm{d}X^2)$ in the frequency domain

- However, this is not a good idea in the presence of noise:
  - The variance of $c_f(\mathrm{d}X^1)c_{-f}(\mathrm{d}X^2)$ due to the noise increases as f increases (cf. Mancino & Sanfelici, 2008)
  - The end effect due to the noise is crucial

- For these reasons
  - We consider "localized" Fourier coefficients and smooth them in the time domain
  - We only use Fourier sine coefficients; they do not suffer from the end effect because $\sin(0) = \sin(2$[redacted]$) = 0$

Construction of a rate-optimal test

- This results in considering *spectral statistics* of Bibinger, Hautsch, Malec & Reiß (2014) (with the lowest frequency):
  - Split $[0\ \ 1]$ into blocks $[kh_n\ \ (k+1)h_n)$ $(k = 0\ \ 1\qquad h_n^{-1} - 1)$
    - h_n is the width of the blocks and chosen so that $h_n^{-1} \in \mathbb{N}$ and $h_n \asymp n^{-\frac{1}{2}}$
  - Define

$$S_k = \sum_{i=1}^{n} (Y_i - Y_{i-1})\, \Phi_k(\bar{t}_i) \qquad \bar{t}_i = \frac{t_{i-1} + t_i}{2}$$

where $\Phi_k(t) = \sin\left(\pi h_n^{-1}(t - kh_n)\right) 1_{[kh_n\ (k+1)h_n)}(t)$

Construction of a rate-optimal test

- To make use of Fourier cosine coefficients, we rely on the same trick as in Bibinger & Winkelmann (2015)
  - We consider the spectral statistics on the shifted blocks $[(k-\frac{1}{2})h_n \;\; (k+\frac{1}{2})h_n)$ as well, i.e. $S_{k-\frac{1}{2}} \; (k = 1 \qquad h_n^{-1} - 1)$
  - Bibinger & Winkelmann (2015) use these statistics to handle jumps in their spectral covariance estimators
- The following formula plays a key role:

$$\Phi_{k-1}(t) - \Phi_k(t) = \cos\left(\pi h_n^{-1}\,(t - (k-1 \;\; 2)\,h_n)\right) 1_{[(k-1)h_n \;\; (k+1)h_n)}(t)$$

- Therefore, noting that $|\ \ | \leq h_n\ \ 2$, we have

$$E\left[\left(S^1_{k-1} - S^1_k\right) S^2_{k-\frac{1}{2}} - S^1_{k-\frac{1}{2}} \left(S^2_{k-1} - S^2_k\right)\right]$$

$$= \frac{\rho}{n} \sum_{(k-\frac{1}{2})h_n \leq \bar{t}_i \ \ (k+\frac{1}{2})h_n} \left\{\cos\left(h_n^{-1} (\bar{t}_i - (k-1\ \ 2) h_n)\right) \sin\left(h_n^{-1} (\bar{t}_i + \quad - (k-1\ \ 2) h_n)\right)\right.$$

$$\left. - \sin\left(h_n^{-1} (\bar{t}_i - (k-1\ \ 2) h_n)\right) \cos\left(h_n^{-1} (\bar{t}_i + \quad - (k-1\ \ 2) h_n)\right)\right\}$$

$$= \rho h_n \sin\left(h_n^{-1} \quad \right)$$

due to the formula $\sin(y - x) = \cos(x)\sin(y) - \sin(x)\cos(y)$

- This motivates us to consider the following moment-type estimator:

$$\Xi_n = \frac{1}{h_n^{-1} - 1} \sum_{k=1}^{h_n^{-1}-1} \left\{\left(S^1_{k-1} - S^1_k\right) S^2_{k-\frac{1}{2}} - S^1_{k-\frac{1}{2}} \left(S^2_{k-1} - S^2_k\right)\right\}$$

Theorem 2

Suppose that $\sqrt{n}h_n \to \kappa$ for some $\kappa \quad 0$. For model (1), we have

$$h_n^{-\frac{3}{2}}\left(\Xi_n - \rho h_n \sin(\blacksquare h_n^{-1} \quad)\right) \xrightarrow{d} N(0 \quad V)$$

as $n \to \infty$, where

$$V = \left\{\left({}^2_1 + \blacksquare^2\kappa^{-2}\Upsilon_1\right)\left({}^2_2 + \blacksquare^2\kappa^{-2}\Upsilon_2\right) - (\quad {}_1 \quad {}_2\rho)^2\right\}$$
$$+ \blacksquare^{-2}\left\{\left({}^2_1 - \blacksquare^2\kappa^{-2}\Upsilon_1\right)\left({}^2_2 - \blacksquare^2\kappa^{-2}\Upsilon_2\right) - (\quad {}_1 \quad {}_2\rho)^2\right\}$$

Construction of a rate-optimal test

- Theorem 2 suggests the test rejecting H_0 if $|T_n^{\mathrm{sp}}| \quad z_{1-\alpha\ 2}$, where $T_n^{\mathrm{sp}} = h_n^{-\frac{3}{2}} \Xi_n \quad \sqrt{V}$
- $h_n^{-\frac{3}{2}} \asymp n^{\frac{3}{4}}$ and $h_n \sin(\blacksquare h_n^{-1} \quad) \asymp \quad$ (because $\quad = o(h_n)$) and we can directly check

$$\limsup_{n\to\infty} \sup_{|c|\le\delta_c} E_{n\ c}\left[\left|h_n^{-\frac{3}{2}}\left(\Xi_n - \rho h_n \sin(\blacksquare h_n^{-1} \quad)\right)\right|^r\right] \quad \infty$$

for all $r \quad 1$ (because Ξ_n is moment-type), so the test based on T_n^{sp} is indeed rate-optimal

Construction of a feasible test

- The test T_n^{sp} is infeasible in practice because V contains the parameters $\;_1 \quad _2\ \rho\ \ \Upsilon_1\ \ \Upsilon_2$ which are usually unknown
- However, a feasible test can be obtained once we construct a consistent estimator for V, and it is an easy task: Set

$$\widehat{\Upsilon}_p^n = -\frac{1}{n}\sum_{i=1}^{n-1}(Y_i^p - Y_{i-1}^p)(Y_{i+1}^p - Y_i^p)$$

$$\widehat{\Sigma}_{pq}^n = \sum_{k=1}^{h_n^{-1}-1}\left(S_k^p S_k^q + S_{k-1\ 2}^p S_{k-1\ 2}^q\right) - \frac{[illegible]^2}{nh_n^2}\widehat{\Upsilon}_p^n 1_{\{p=q\}}$$

Construction of a feasible test

- We have $\widehat{\Upsilon}^n_p \to^p \Upsilon_p$, $\widehat{\Sigma}^n_{pp} \to^p \;\; {}^2_p$ and $\widehat{\Sigma}^n_{12} \to^p \;\; {}_1 \;\; {}_2 \rho$

  $\Longrightarrow$ Setting $\widehat{\Sigma}^n_{p\,\pm} = \widehat{\Sigma}^n_{pp} \pm (\blacksquare^2 \;\; n h_n^2)\widehat{\Upsilon}^n_p$ and

$$\widehat{V}^n = \widehat{\Sigma}^n_{1\,+}\widehat{\Sigma}^n_{2\,+} + \blacksquare^{-2}\widehat{\Sigma}^n_{1\,-}\widehat{\Sigma}^n_{2\,-} - (1 + \blacksquare^{-2})\left(\widehat{\Sigma}^n_{12}\right)^2$$

  we have $\widehat{V}^n \to^p V$

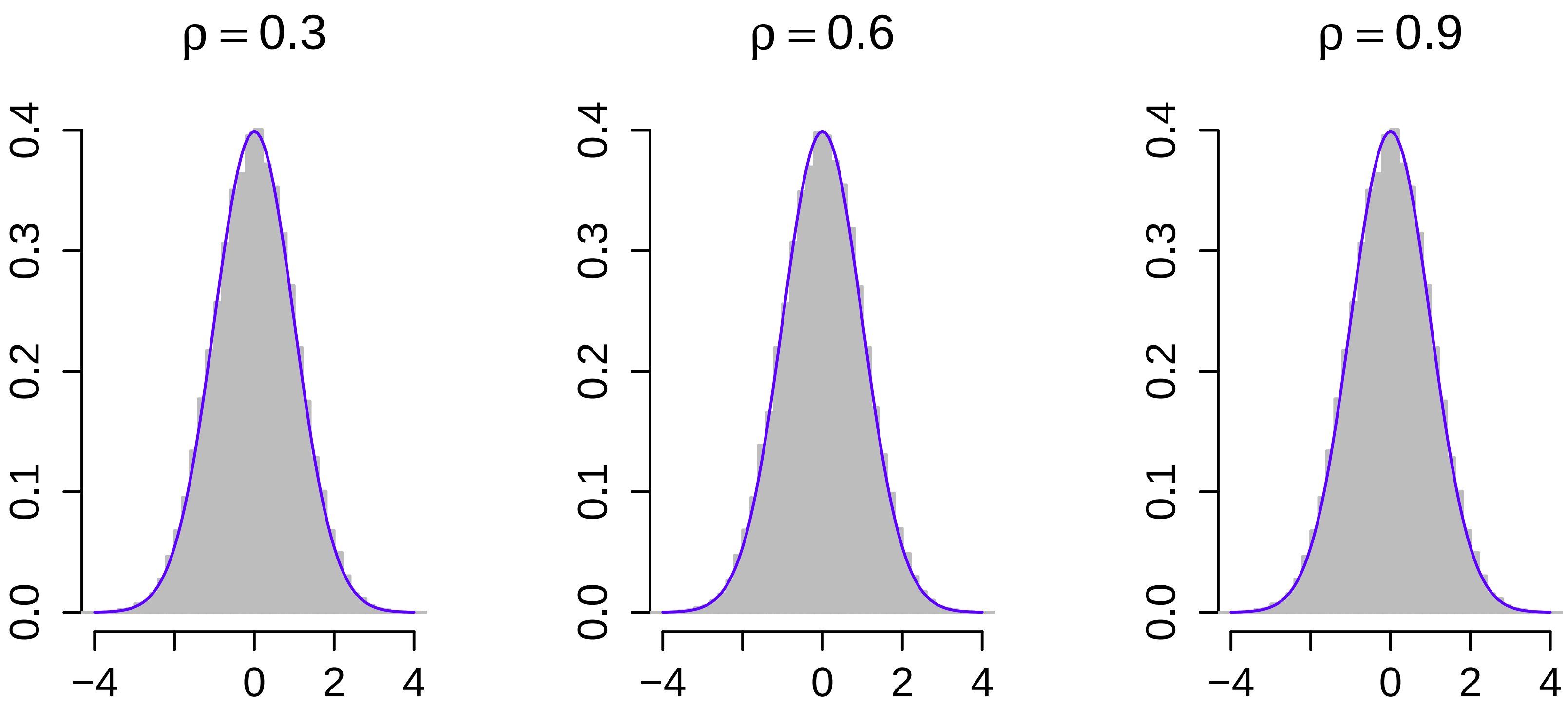
- Consequently, we obtain a feasible test statistic

$$\widehat{T}^{\text{sp}}_n = h_n^{-\frac{3}{2}} \frac{\Xi_n}{\left(\widehat{V}^n\right)^{1 \;\; 2}}$$

Simulation study

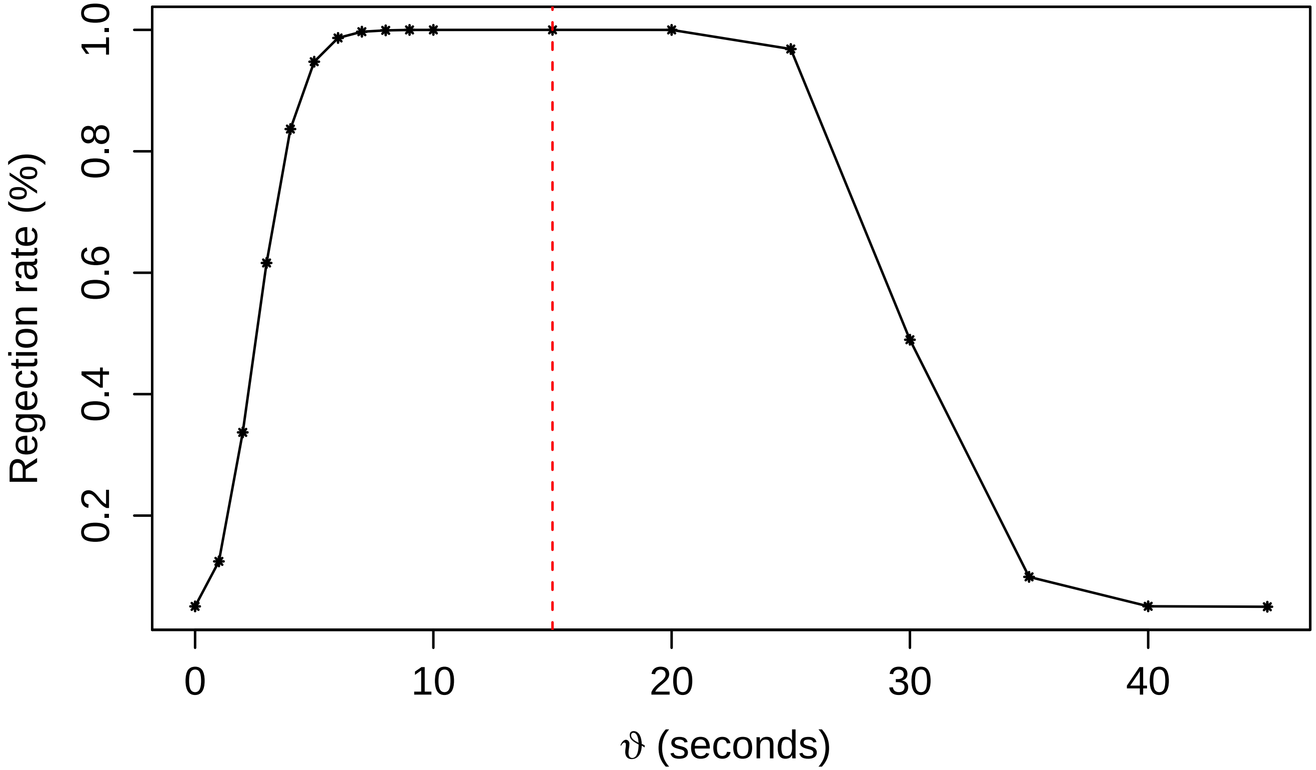
- We set $\ _p = 1$, $\Upsilon_p = 0\ 001$ for $p = 1\ \ 2$ and $\rho \in \{0\ 3\ \ 0\ 6\ \ 0\ 9\}$
  - The noise variance is 0.1% of the quadratic variation, reflecting the empirical finding of Hansen & Lunde (2006)
- $n = 3\ \ 600$
- We regard $\frac{1}{n}$ as 1 second, so $[0\ \ 1]$ corresponds to 1 hour
- $\ = l\ \ n$ and $l = 0\ \ 1 \qquad 10\ \ 15\ \ 20 \qquad 45$
- $h_n = 30\ \ n$; note that the consistency of the test is not ensured at the lags higher than $\ = h_n\ \ 2 = 15\ \ n$

Figure 1: Histograms of $\widehat{T}_n^{\mathrm{sp}}$ under H_0



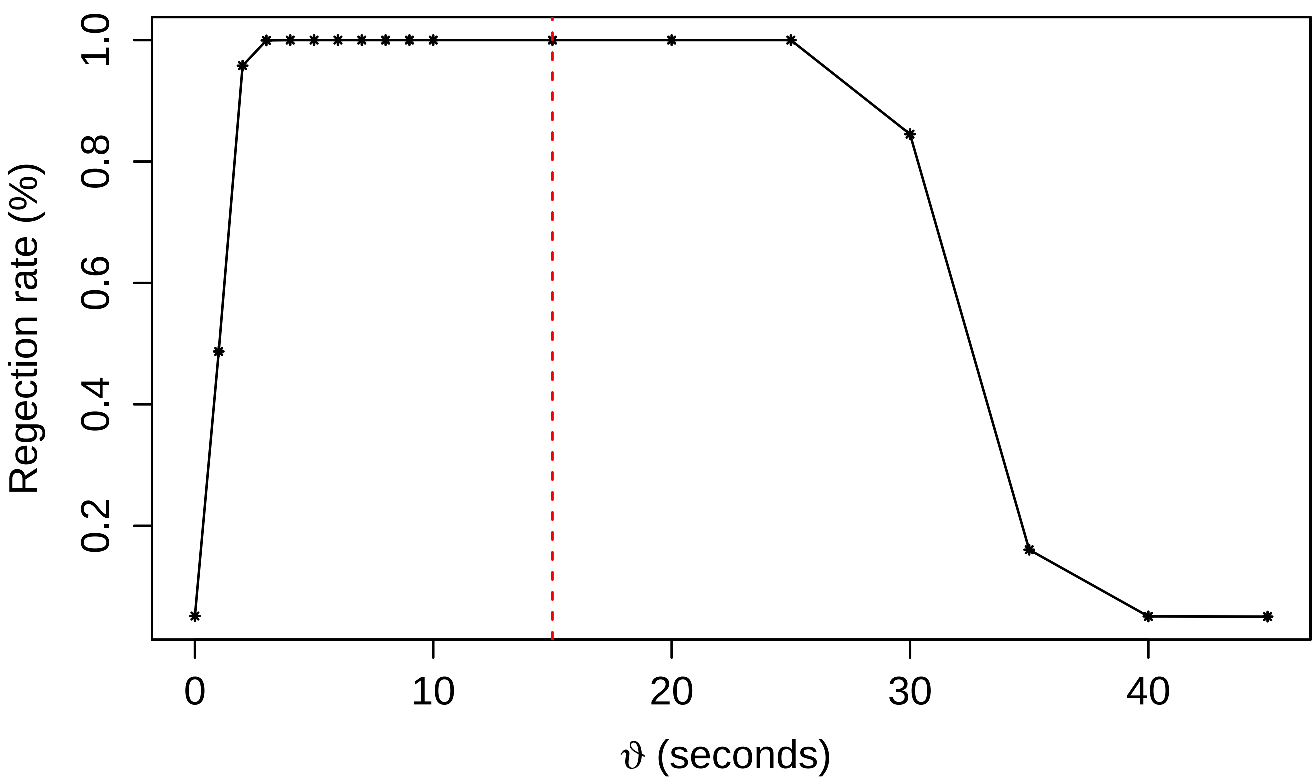
Note. Monte Carlo distribution of $\widehat{T}_n^{\mathrm{sp}}$ under H_0 based on 50,000 repetitions (grey). Blue solid lines denote the $N(0, 1)$ density.

Figure 2: Rejection rate of H_0 at the 5% level ($\rho = 0\ 6$)



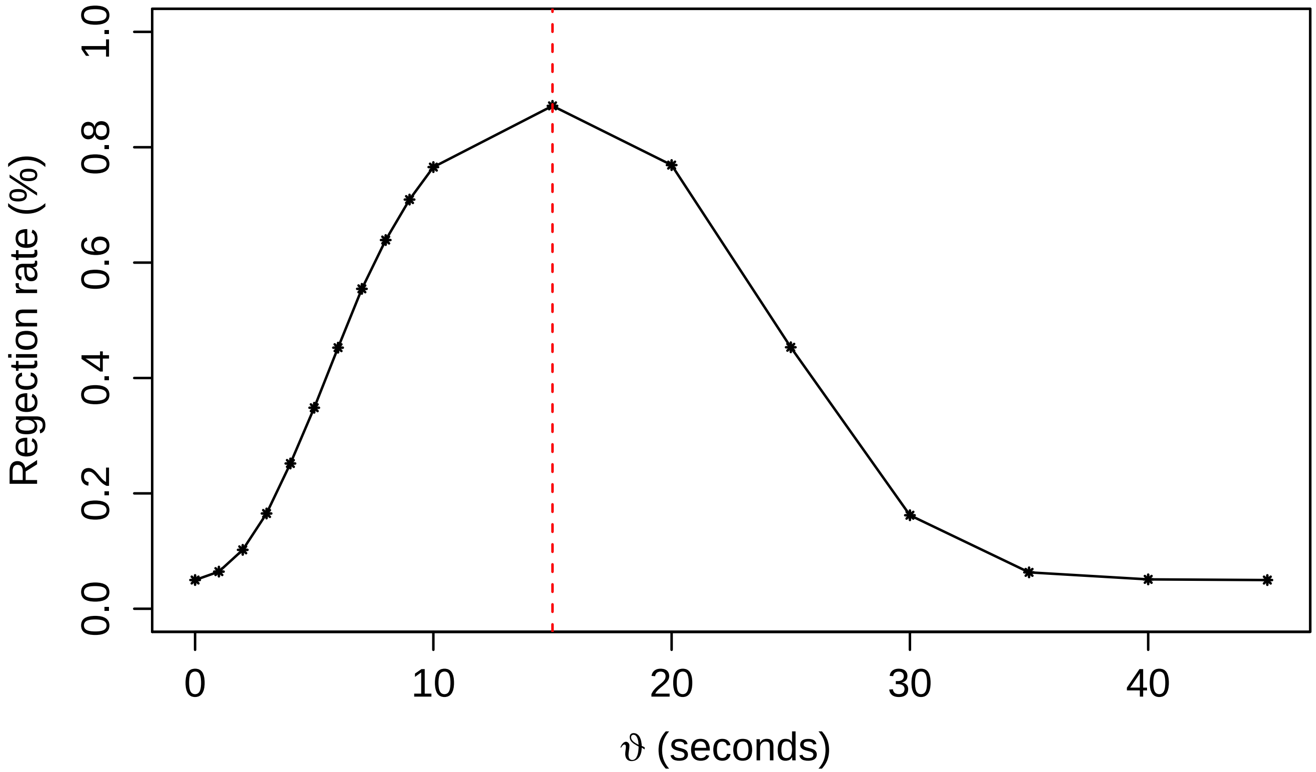
Note. Monte Carlo empirical rejection rate of H_0 at the 5% level based on 50,000 repetitions ($\rho = 0\ 6$). Red dash line denotes $\ = h_n\ 2$.

Figure 3: Rejection rate of H_0 at the 5% level ($\rho = 0\ 9$)



Note. Monte Carlo empirical rejection rate of H_0 at the 5% level based on 50,000 repetitions ($\rho = 0\ 9$). Red dash line denotes $\quad = h_n\ \ 2$.

Figure 4: Rejection rate of H_0 at the 5% level ($\rho = 0\ 3$)



Note. Monte Carlo empirical rejection rate of H_0 at the 5% level based on 50,000 repetitions ($\rho = 0\ 3$). Red dash line denotes $\ = h_n\ 2$.

Conclusions

- Contributions of this study
  - For the Hoffmann-Rosenbaum-Yoshida model of lead-lag effects, lower bounds of detectable lags' rate have been provided both in the non-noisy case and the noisy case
  - In the noisy case, a simple feasible test that attains the optimal rate is proposed
- Future works
  - Extension of the model: stochastic volatility and non-synchronous observations (probably routine)
  - More general model of lags (e.g. time varying one)

References

Bacry, E., Delattre, S., Hoffmann, M. & Muzy, J. (2013). Modelling microstructure noise with mutually exciting point processes. *Quant. Finance* **13**, 65–77.

Bibinger, M., Hautsch, N., Malec, P. & Reiß, M. (2014). Estimating the quadratic covariation matrix from noisy observations: local method of moments and efficiency. *Ann. Statist.* **42**, 80–114.

Bibinger, M. & Winkelmann, L. (2015). Econometrics of co-jumps in high-frequency data with noise. *J. Econometrics* **184**, 361–378.

Campbell, J. Y., Lo, A. W. & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton University Press.

Christensen, K., Kinnebrock, S. & Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *J. Econometrics* **159**, 116–133.

Gloter, A. & Jacod, J. (2001). Diffusions with measurement errors. I. Local asymptotic normality. *ESAIM Probab. Stat.* **5**, 225–242.

Granger, C. & Hatanaka, M. (1964). *Spectral analysis of economic time series*. Princeton University Press.

Griliches, Z. (1967). Distributed lags: A survey. *Econometrica* **35**, 16–49.

Hansen, P. R. & Lunde, A. (2006). Realized variance and market microstructure noise. *J. Bus. Econom. Statist.* **24**, 127–161.

Hoffmann, M., Rosenbaum, M. & Yoshida, N. (2013). Estimation of the lead-lag parameter from non-synchronous data. *Bernoulli* **19**, 426–461.

Huth, N. & Abergel, F. (2014). High frequency lead/lag relationships — empirical facts. *Journal of Empirical Finance* **26**, 41–58.

Ingster, Y. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I, II, III. *Math. Methods Statist.* **2**, 85–114; 171–189; 249–268.

Malliavin, P. & Mancino, M. E. (2009). A Fourier transform method for nonparametric estimation of multivariate volatility. *Ann. Statist.* **37**, 1983–2010.

Mancino, M. E. & Sanfelici, S. (2008). Robustness of Fourier estimator of integrated volatility in the presence of microstructure noise. *Comput. Statist. Data Anal.* **52**, 2966–2989.

Podolskij, M. & Vetter, M. (2009). Bipower-type estimation in a noisy diffusion setting. *Stochastic Process. Appl.* **119**, 2803–2831.

Robert, C. Y. & Rosenbaum, M. (2010). On the limiting spectral distribution of the covariance matrices of time-lagged processes. *J. Multivariate Anal.* **101**, 2434–2451.

Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24**, 2477–2498.

Vetter, M. & Dette, H. (2012). Model checks for the volatility under microstructure noise. *Bernoulli* **18**, 1421–1447.