Multi-Scale Analysis of Lead-Lag Relationships in High-Frequency Financial Markets [1]

Yuta Koike

University of Tokyo, CREST JST

December 1, 2020

The LiU Seminar Series in Statistics and Mathematical Statistics

[1] Joint work with Takaki Hayashi (Keio University)

1. Background
2. Model
3. Estimation
4. Empirical application
5. Conclusions

Background

- **Lead-lag relationship**
  - Two time series are cross-correlated with each other at certain lags; "leader" and "lagger"
- Lead-lag relationships may occur perhaps because new information is absorbed into each security at different speeds
  - Across different assets
  - Across different trading venues
- **Ex.**: Stock index vs index futures (e.g. Kawaller *et al.*, 1987)
  - A stock index consists of many individual stocks; it may be lagging behind the index futures

Lead-lag analysis with high-frequency data

- Timestamps are very important in high-frequency data, necessarily to be modeled
  - Discretely observed continuous-time processes are appropriate
- Price series based analysis
  - continuous semimartingale based model, utilizing the Hayashi-Yoshida estimator (Hayashi and Yoshida, 05)
    - ⋆ Hoffman, Rosenbaum and Yoshida (13), Huth and Abergel (14)
  - multivariate Hawkes processes based model
    - ⋆ Bacry, Delattre, Hoffmann Muzy (11), Da Fonseca and Zaatour (14)
- Timestamp based analysis
  - based on the counts of the co-occurrent "events"
    - ⋆ Dobrev and Schaumburg (16)

Background: HRY model

Hoffmann, Rosenbaum & Yoshida (2013)

- Suppose that the log-price processes of two assets are given by

$$\begin{cases} X_t^1 = \sigma_1 W_t^1, \\ X_t^2 = \sigma_2 \rho W_{t-\vartheta}^1 + \sigma_2 \sqrt{1-\rho^2} W_{t-\vartheta}^2, \end{cases} \tag{1}$$

  where
  - $\sigma_1, \sigma_2 > 0$, $\rho \in [-1, 1]$ and $\vartheta \in (-\delta, \delta)$ are constants
  - W^1, W^2 are independent Brownian motions
- $0 \leq t_1^\nu < t_2^\nu < \cdots < t_{n_\nu}^\nu \leq T$: observation times for X^ν
  - could be different across two assets (non-synchronous observations)
- Idea to estimate the time-lag ϑ
  - The returns of $(X_{t_i^1}^1)_{i=0}^{n_1}$ and $(X_{t_j^2}^2)_{j=0}^{n_2}$ are (significantly) cross-correlated only at the lag ϑ
  - Maximizer of their (empirical) CCF will be a good estimator for ϑ

Background: HRY estimator

- How to compute the CCF ?
  $\Longrightarrow$ time-lagged version of the HY estimator:

$$U^{HRY}(\theta) = \sum_{i,j} \Delta_i X^1 \Delta_j X^2 1_{\{(t^1_{i-1}, t^1_i] \cap (t^2_{j-1} - \theta, t^2_j - \theta] \neq \emptyset\}},$$

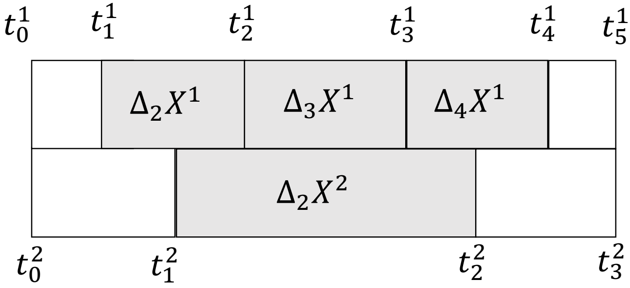
  where $\Delta_i X^\nu = X^\nu_{t^\nu_i} - X^\nu_{t^\nu_{i-1}}$ for $\nu = 1, 2$
- Hoffmann *et al.* (2013) have shown that

$$\widehat{\theta}^{HRY} = \arg\max_{\theta \in \mathcal{G}} |U^{HRY}(\theta)|$$

  is a consistent estimator for ϑ under some regularity conditions while one appropriately takes the finite set $\mathcal{G} \subset (-\delta, \delta)$
  - The method works for more general diffusion processes
  - The R package **yuima** contains the function `llag` to implement $\widehat{\theta}^{HRY}$

Background: HRY estimator

Figure 1: The Hayashi-Yoshida method: We sum up cross-products of returns with overlapping observation intervals



Background: DS estimator

Dobrev & Schaumburg (2016)

- For $\nu = 1, 2$ and $t \geq 0$, we set

$$I_t^\nu = \begin{cases} 1 & \text{if the } \nu\text{-th asset is observed at the time } t, \\ 0 & \text{otherwise} \end{cases}$$

- We count the co-occurrent observations with the time lag $\theta \in \mathbb{R}$ by

$$U^{DS}(\theta) = \frac{1}{\min\{n_1, n_2\}} \sum_{k=1}^{\lfloor T/\tau_N \rfloor} I^1_{k\tau_N} I^2_{k\tau_N+\theta},$$

where τ_N is the finest time resolution in analysis (0.1ms in our case)
- The DS estimator $\widehat{\theta}^{DS}$ is defined as a maximizer of $U^{DS}(\theta)$ over a grid $\mathcal{G}$:

$$\widehat{\theta}^{DS} = \arg\max_{\theta \in \mathcal{G}} U^{DS}(\theta).$$

Background: An empirical illustration

Lead-lag analysis of the NASDAQ-100 assets: NASDAQ vs BATS

- There are 13 major stock exchanges in the U.S. stock market, and one can send orders to any exchanges
  - A single asset may have different prices at each exchange
    $\Longrightarrow$ A lead-lag relationship could appear between different exchanges
- We examine lead-lag relationships between the NASDAQ and BATS exchanges for each component stock of NASDAQ-100 in 2015 (totally 108 assets)
  - Period: All the trading days in August, 2015 (totally 21 days)
    - Between 9:45 and 15:45 (the first and the last 15 min are discarded to exclude abnormal behaviors at the opening and closing)
  - Data source: Best quote data from the Daily TAQ Database
    - The precision of timestamps is in micro-seconds, but we set the finest time resolution in analysis as $\tau_N = 0.1$ms due to the reasons explained later

Background: An empirical illustration

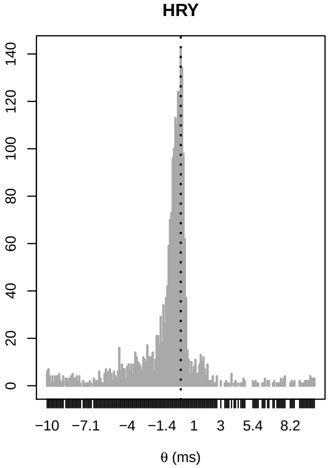
Lead-lag analysis of the NASDAQ-100 assets: NASDAQ vs BATS

- Best quote data contains the following information:
  - Best ask price p_a and its volume v_a (the lowest price accepted by a seller)
  - Best bid price p_b and its volume v_b (the highest price accepted by a buyer)
- We construct price processes from these data by computing the so-called micro-price (cf. Gatheral & Oomen (2010)):

$$q_v := \frac{p_a/v_a + p_b/v_b}{1/v_a + 1/v_b} = \frac{v_b p_a + v_a p_b}{v_b + v_a}$$

- We set $\mathcal{G} = \{-10.0\text{ms}, -9.9\text{ms}, \dots, 9.9\text{ms}, 10.0\text{ms}\}$ and compute the HRY and DS estimators for each asset on each trading day
  $\Rightarrow$ We get totally $21 \times 108 = 2268$ estimates for these two estimators

Figure 2: Histograms of the daily lead-lag time estimates for the NASDAQ-100 assets



$\theta > 0$ indicates that the NASDAQ leads the BATS.

Background: An empirical illustration

Lead-lag analysis of the NASDAQ-100 assets: NASDAQ vs BATS

- Most DS estimates concentrate at $\theta = +0.3$ms, suggesting that the NASDAQ consistently leads the BATS with the lag 0.3ms.
- In contrast, the HRY estimates are negatively skewed, suggesting the BATS tends to lead the NASDAQ
- As explained below, the DS estimates 0.3ms would come from a geographical reason:
  - Transit time btw the NASDAQ and BATS ($\approx$ 0.1ms) + Reporting latency from the BATS ($\approx$ 0.2ms)

Background: An empirical illustration

A geographical consideration

- Dobrev & Schaumburg (2016) argued that the DS estimator captures the transit time of information btw two venues in cross-market analysis
- In our situation,
  - Servers of the NASDAQ @ Carteret, NJ
  - Servers of the BATS @ Secaucus, NJ
  - The minimum transit time btw Carteret and Secaucus (in the speed of light) $\approx$ 0.09ms (Tivnan *et al.*, 2020, Table 2)
- Our DS estimate $=$ 0.3ms $\Rightarrow$ Where does the extra 0.2ms come from ?

NMS Propagation Delay Estimates				
	Carteret-Mahwah	Mahwah-Secaucus	Carteret-Secaucus	Secaucus-Weehawken
Straight-line Distance	34.55 mi	21.31 mi	16.22 mi	2.56 mi
	55.6 km	34.3 km	26.1 km	4.12 km
Light speed, one-way	185.75 μs	114.57 μs	87.2 μs	13.76 μs
Light speed, two-way	371.5 μs	229.14 μs	174.4 μs	27.52 μs
Fiber, one-way	272.44 μs	168.07 μs	127.89 μs	20.19 μs
Fiber, two-way	544.88 μs	336.14 μs	255.78 μs	40.38 μs
Hybrid laser, one-way	-	-	94.5 μs	-
Hybrid laser, two-way	-	-	189 μs	-

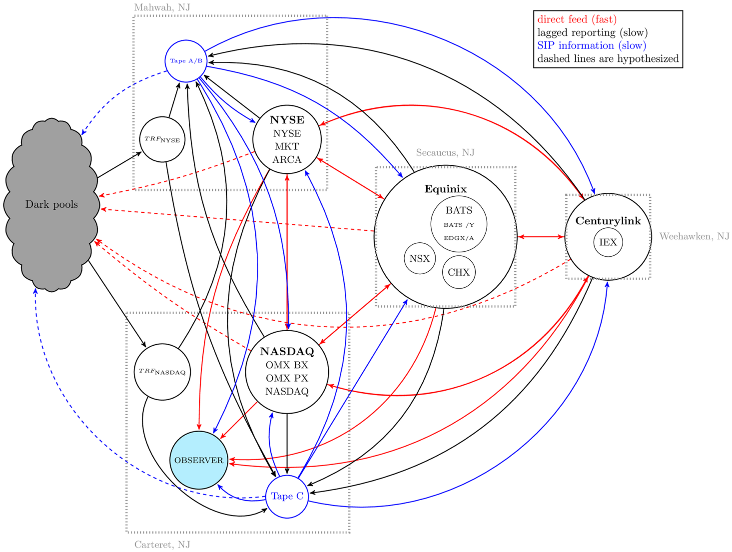
https://doi.org/10.1371/journal.pone.0226968.t002

Source: (Tivnan *et al.*, 2020, Table 2)

Background: An empirical illustration

A geographical consideration

- In U.S. stock market, all orders are consolidated into a single data feed by *Securities Information Processors (SIPs)*
- The Daily TAQ database provides timestamps when the corresponding orders are processed by SIPs rather than exchanges
  $\Rightarrow$ **We need to take account of time-lags to send orders from exchanges to SIPs**
- For NASDAQ-listed stocks, the corresponding SIP is located at Carteret, yielding around 0.2ms reporting latencies from the BATS compared with the NASDAQ (Bartlett & McCrary, 2019)



Sourse: (Tivnan *et al.*, 2020, Figure 1)

Background: An empirical illustration

Fact or Failure ?

- These considerations would suggest the relevance of the DS estimator
- If the time-lag is caused by a purely geographical reason, we may naturally expect $\widehat{\theta}^{HRY}$ should be close to 0.3ms as above
  $\Longrightarrow$ Does the HRY model fail to capture right relationships?
- We conjecture that the "failure" of the HRY model is due to the *market heterogeneity*
  - **"Heterogenous market hypothesis"** (Müller *et al.*, 1997): Market participants act with different time scales
  - Different lead-lags can coexist at different time scales
  - The HRY model would capture lead-lag relationships coming from different time scales
  - Wall street is closer to Secaucus than Carteret, so the BATS receives orders of "slow" traders faster than the NASDAQ

Source:
https://www.nytimes.com/2013/05/14/technology/north-jersey-data-center-industry-blurs-utility-real-estate-boundaries.html

Our contribution

- We propose a model taking account of "heterogeneity" of the market
  - Modeling with multiple time scales
    $\Longrightarrow$ Wavelets !! (cf. Gençay *et al.*, 2002)
- The existing literature on applications of wavelet to lead-lag analysis is based on *discrete-time* modeling (mainly established in Whitcher *et al.* (1999, 2000) and Serroukh & Walden (2000a,b))
- **Contribution of this work**
  - Providing a modeling framework validating wavelet analysis for investigating lead-lag relationships with multiple time scales in a *continuos-time* setting
  - Proposing an estimation procedure for the lead-lag parameters on a scale-by-scale basis

Model

- $\boxed{\text{Idea}}$ Characterize the lead-lag relationship of two BMs in the frequency domain
  - The theoretical CCF of dB^1 and dB^2 is not a proper function
  - The cross-spectral density of dB^1 and dB^2 always exists as a proper function (Hayashi and K. (2018), Proposition 2)
- The HRY model (1) has the cross-spectral density given by

$$f(\lambda) = \sigma_1 \sigma_2 \rho e^{-\sqrt{-1}\lambda\vartheta}, \qquad \lambda \in \mathbb{R}$$

- We split the frequency domain into "octave" bands:

$$\Lambda_j := [-2^{j+1}\pi, -2^j\pi) \cup (2^j\pi, 2^{j+1}\pi], \qquad j = 0, 1, \dots$$

Model

- From the spectral/wavelet analysis perspective, Λ_j is regarded as the component corresponding to the time scale $[2^{-j}, 2^{-j+1})$

  $\Rightarrow$ We wish to consider the cross-spectral density of the form

$$f_N(\lambda) = \sum_{j=1}^{N+1} R_j e^{-\sqrt{-1}\lambda\theta_j} 1_{\Lambda_{N-j+1}}(\lambda), \tag{2}$$

  where
  - N is the finest resolution level ($j = 1$ corresponds to this level)
  - R_j is a non-zero number (the correlation at the frequency band Λ_{N-j+1})
  - θ_j is the lead-lag time parameter at the frequency band Λ_{N-j+1}
  - Taking T appropriately, we let $\tau_N := 2^{-N+1}$ correspond to the finest time resolution in analysis

Model

- In fact, we can construct a bivariate Gaussian process $B_t = (B_t^1, B_t^2)$ with stationary increments such that
  (i) The respective marginal processes B_t^1 and B_t^2 are standard BMs
  (ii) B_t has the cross-spectral density given by (2)
    - See Hayashi and K. (2018, Proposition 2)
- We suppose that the log-price processes of two assets are given by[2]

$$X_t^1 = \sigma_1 B_t^1, \qquad X_t^2 = \sigma_2 B_t^2$$

- We wish to estimate the time-lag parameters θ_j from the observation data $(X^1_{t^1_i})_{i=0}^{n_1}$ and $(X^2_{t^2_j})_{j=0}^{n_2}$

[2] Extension to the stochastic volatility case is possible; see our working paper **arXiv:1708.03992v4**

Estimation: Wavelet decomposition of the CCF

- Let $U^N(\theta)$ be the inverse Fourier transform of $f_N(\lambda)$
  - $U^N(\theta)$ corresponds to the theoretical CCF of dB^1 and dB^2
- $U^N(\theta)$ admits the following decomposition:

$$U^N(\theta) = \sum_{j=1}^{N+1} \sigma_1\sigma_2 R_j 2^{N-j+1}\psi^{LP}(2^{N-j+1}(\theta-\theta_j)),$$

where

$$\psi^{LP}(s) := (\pi s)^{-1}(\sin(2\pi s) - \sin(\pi s))$$

is known as the *Littlewood-Paley wavelet*

Estimation: Wavelet decomposition of the CCF

- Using the property of the LP wavelet, we have

$$\begin{aligned}\rho_{(j)}(\theta) := &\int_{-\infty}^{\infty} U^N(\theta - s)\psi^{LP}(2^{N-j+1}s)ds \\ &= \sigma_1\sigma_2 R_j \psi(2^{N-j+1}(\theta - \theta_j))\end{aligned}$$

  - We shall regard $\rho_{(j)}(\theta)$ as a "CCF at the level j"
  - θ_j is the unique maximizer of $|\rho_{(j)}(\theta)|$ as long as $R_j \neq 0$
- The expression of $\rho_{(j)}(\theta)$ naturally suggests the following estimator:

$$\widehat{\rho}_{(j)}(\theta) = \sum_{l=-L_j+1}^{L_j-1} U^{HRY}(\theta - l\tau_N)\Psi_j(l),$$

where $\Psi_j(l)$ is an approximation of $\psi^{LP}(2^{N-j+1}l\tau_N)$

Estimation: Approximation of LP wavelets

- We may directly use $\Psi_j(l) = \psi^{LP}(2^{N-j+1} l \tau_N)$, but there is a mathematically preferable alternative
- The Fourier inversion formula yields

$$\psi^{LP}(2^{N-j+1} l \tau_N) = 2^j \int_{-\pi}^{\pi} e^{\sqrt{-1} l \lambda} 1_{\Lambda_{-j}}(\lambda) d\lambda$$

  - The transfer function of $(\psi^{LP}(2^{N-j+1} l \tau_N))_{l \in \mathbb{Z}}$ is $2^j 1_{\Lambda_{-j}}(\lambda)$
  - $\Psi_j(l)$ well approximates $\psi^{LP}(2^{N-j+1} l \tau_N)$ if the transfer function of $(\Psi_j(l))_{l=-L_j+1}^{L_j-1}$ well approximates $2^j 1_{\Lambda_{-j}}(\lambda)$
- We utilize Daubechies' wavelet filters to construct such $\Psi_j(l)$'s

Estimation: Approximation of LP wavelets

- Let $h_{j,0}, h_{j,1}, \dots, h_{j,L_j-1}$ be Daubechies' wavelet filters with length L at the level j ($L_j = (2^j - 1)(L-1) + 1$)
  - $L = 2$ corresponds to the Haar wavelet filters
- The power transfer function $H_{j,L}(\lambda) = |\sum_{p=0}^{L_j-1} h_{j,p} e^{-\sqrt{-1}\lambda p}|^2$ well approximates $2^j 1_{\Lambda_{-j}}(\lambda)$ as $L \to \infty$ (Lai, 1995)
- This suggests us to set

$$\Psi_j(l) = \sum_{p=0}^{L_j-1-|l|} h_{j,p} h_{j,p+|l|}, \qquad l = 0, \pm 1, \dots, \pm(L_j - 1) \tag{3}$$

  - This is known as the *autocorrelation wavelets* (cf. Nason *et al.*, 2000)

Estimation

- Since θ_j is the unique maximizer of $|\rho_{(j)}(\theta)|$ (if $R_j \neq 0$), we naturally estimate it by

$$\widehat{\theta}_j := \arg\max_{\theta \in \mathcal{G}_j^N} \left|\widehat{\rho}_{(j)}(\theta)\right|$$

- To avoid boundary issues, we take

$$\mathcal{G}_j^N = \{l\tau_N : l \in \mathbb{Z}, |l\tau_N| < \delta - L_j\tau_N\}$$

- The following result ensures the consistency of our estimators

Theorem 1 (Hayashi and K. (2020), Theorem 2)

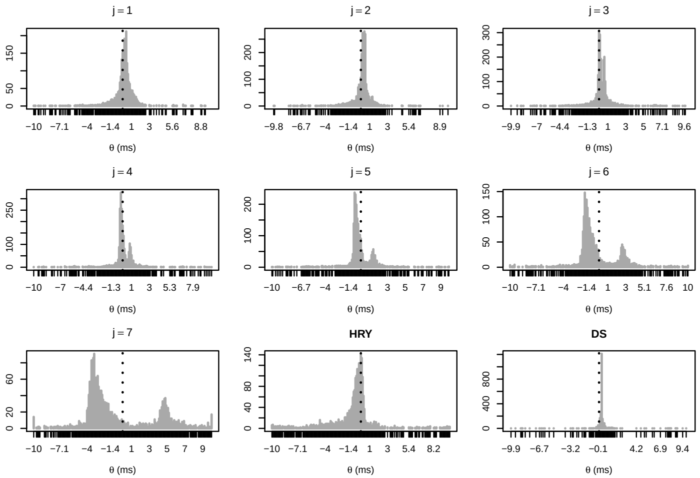
Suppose that $L \to \infty$ and $\tau_N L \log L \to 0$ as $N \to \infty$. Under some regularity conditions on the observation times, we have $\widehat{\theta}_j \to^p \theta_j$ as $N \to \infty$ for every j with $R_j \neq 0$.

Empirical application

Lead-lag analysis of the NASDAQ-100 assets: NASDAQ vs BATS

- Cross-market, single-asset analysis
- Venues: NASDAQ and BATS
- Micro price (inverse-volume weighted mid-quote)
- Stocks: Component stocks of NASDAQ-100 in 2015
- Source: Daily TAQ Database
- The data are recorded in micro-secs, but we set $\tau_N = 0.1$ms due to a clock synchronization issue
- Period: All the trading days in August, 2015
- Between 9:45 and 15:45 (the first and the last 15 min are discarded)
- Search grid: $\mathcal{G}_j^N = \{-10.0\text{ms}, -9.9\text{ms}, \dots, 9.9\text{ms}, 10.0\text{ms}\}$
- $L = 20$ (length of Daubechies' wavelet filters)

Figure 3: Histograms of the daily lead-lag time estimates for the NASDAQ-100 assets



$\theta > 0$ indicates that the NASDAQ leads the BATS.

Empirical application

- We find that
  - the estimates of $\widehat{\theta}_j$ at the levels $j = 1, 2, 3$ have sharp peaks at small positive values
  - the estimates of $\widehat{\theta}_j$ at the levels $j = 4, 5, 6, 7$ have two peaks located at positive and negative values, respectively
- These observations suggest that
  - the estimates of $\widehat{\theta}_j$ at the finer levels might be related to those of $\widehat{\theta}^{DS}$ (corresponding to the time scales between 0.1ms and 0.8ms)
  - the negative estimates of $\widehat{\theta}_j$ at the coarser levels $j = 4, 5, 6, 7$ might have some links with those of $\widehat{\theta}^{HRY}$ (corresponding to the time scales between 0.8ms and 12.8ms)
- See our working paper **arXiv:1708.03992v4** for more detailed analysis

Conclusions

- We have introduced a new framework to model and estimate multiple lead-lag relationships in high-frequency data on a scale-by-scale basis
- In the empirical application, we have identified two types of lead-lag relationships at finer and (relatively) coarser time scales, respectively

- This talk is based on the following two papers:
  - T. Hayashi, Y. Koike (2018). "Wavelet-based methods for high-frequency lead-lag analysis", *SIAM J. Financial Math.* **9**, 1208 – 1248.
  - T. Hayashi, Y. Koike (2020). "Multi-scale analysis of lead-lag relationships in high-frequency financial markets", Working paper. Available at `https://arxiv.org/abs/1708.03992v4`
- The R package **yuima** contains the function `wllag` to implement $\widehat{\theta}_j$

References I

Bartlett, R. P. & McCrary, J. (2019). How rigged are stock markets? Evidence from microsecond timestamps. *Journal of Financial Markets* **45**, 37–60.

Dobrev, D. & Schaumburg, E. (2016). High-frequency cross-market trading: Model free measurement and applications. Working paper.

Gatheral, J. & Oomen, R. C. (2010). Zero-intelligence realized variance estimation. *Finance Stoch.* **14**, 249–283.

Gençay, R., Selçuk, F. & Whitcher, B. (2002). *An introduction to wavelets and other filtering methods in finance and economics*. Academic Press.

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

Kawaller, I. G., Koch, P. D. & Koch, T. W. (1987). The temporal price relationship between S&P 500 futures and the S&P 500 index. *Journal of Finance* **42**, 1309–1329.

Lai, M.-J. (1995). On the digital filter associated with Daubechies' wavelets. *IEEE Trans. Signal Process.* **43**, 2203–2205.

References II

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V. & von Weizsäcker, J. E. (1997). Volatilities of different time resolutions — analyzing the dynamics of market components. *Journal of Empirical Finance* **4**, 213–239.

Nason, G. P., von Sachs, R. & Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62**, 271–292.

Serroukh, A. & Walden, A. (2000a). Wavelet scale analysis of bivariate time series i: motivation and estimation. *J. Nonparametr. Stat.* **13**, 1–36.

Serroukh, A. & Walden, A. (2000b). Wavelet scale analysisof bivariate time series ii:statistical properties for linear processes. *J. Nonparametr. Stat.* **13**, 37–56.

Tivnan, B. F., Dewhurst, D. R., Van Oort, C. M., Ring, J. H., Gray, T. J., Tivnan, B. F., Koehler, M. T. K., McMahon, M. T., Slater, D. M., Veneman, J. G. & Danforth, C. M. (2020). Fragmentation and inefficiencies in US equity markets: Evidence from the Dow 30. *PLoS ONE* **15**, e0226968.

Whitcher, B., Guttorp, P. & Percival, D. B. (1999). Mathematical background for wavelet estimators of cross-covariance and cross-correlation. Technical report 038, NRCSE.

References III

Whitcher, B., Guttorp, P. & Percival, D. B. (2000). Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research* **105**, 14941–14962.