Hidden leaders: Identifying high-frequency lead-lag structures in a multivariate price formation framework

Fulvio Corsi
Ca' Foscari University of Venice; City University of London

Joint work with:
Giuseppe Buccheri (Scuola Normale Superiore, Pisa)
Stefano Peluso (Catholic University of Milan)

Motivation and objectives

Motivations: 3 stylized facts on HF autocorrelation structure,

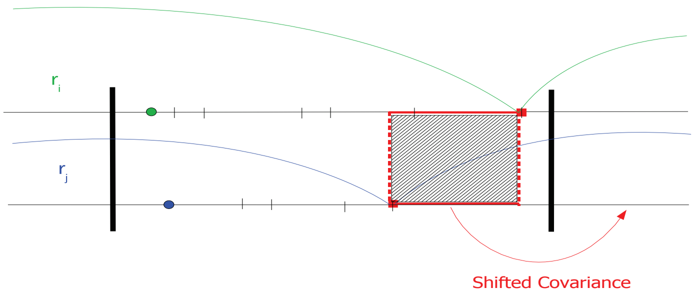
1. Negative 1^{st}-order autocorrelation $\rightarrow$ Roll (1984)
2. Positive higher-order autocorrelation $\rightarrow$ Hasbrouck & Ho (1987)
3. Cross Lead-lag correlation (De Jong & Nijman 1997 Chiao, Hung & Lee 2004, Huth & Abergel 2012) $\rightarrow$ **???**
   $\Rightarrow$ We propose multivariate extension of Hasbrouck & Ho (1987)

Goals of the paper:

- describe lead-lag effects within a theoretical framework which extends well established models proposed in the market microstructure literature
- provide estimation procedure for lead-lag correlations
- test the presence of "true" lead-lag effects
- provide estimator of the Integrated Covariance of efficient price robust to:
  - Market microstructure noise
  - Asynchronous trading
  - Lead-lag dependence
  - And p.d. by construction

Asynchronicity and lead-lag dependence

non-synchronous trading $\Rightarrow$ *shifting portion of contemporaneous covariance*
$\Rightarrow$ *spurious lead-lag dependence*



Microstructure foundations in discrete time

Let $P_t \in \mathbb{R}^d$ be a vector of intraday efficient log-prices $dP_t = \sigma dW_t$

Following standard market microstructure literature on partial price adjustment (Hasbrouck & Ho 1987, Amihud & Mendelson 1987, Damodaran 1992), we rewrite the efficient price process in discrete time

$$P_t = P_{t-1} + u_t, \qquad u_t \sim \text{NID}(0, \Sigma) \tag{1}$$

and define X_t the vector of the latent prices with lagged adjustment

$$X_t = X_{t-1} + \Psi(P_t - X_{t-1}) \tag{2}$$

The matrix Ψ is the speed of market price adjustment

If $\Psi = \mathbb{I}$, then $X_t = P_t$, i.e. instantaneous price adjustment

(1) and (2) imply:

$$\Delta X_{t+1} = (\mathbb{I} - \Psi)\Delta X_t + \Psi u_t \tag{3}$$

a VAR(1) process which we rewrite as

$$\Delta X_{t+1} = F\Delta X_t + \eta_t, \qquad \eta_t \sim \text{NID}(0, Q) \tag{4}$$

with $F = \mathbb{I} - \Psi$ and $Q = \Psi\Sigma\Psi'$

State-space representation

With $Y_t \in \mathbb{R}^d$ the observed log-prices contaminated with MM errors we have:

$$Y_t = X_t + \epsilon_t, \qquad \epsilon_t \sim \mathrm{NID}(0, H) \tag{5}$$
$$\Delta X_{t+1} = F \Delta X_t + \eta_t, \quad \eta_t \sim \mathrm{NID}(0, Q) \tag{6}$$

with H the diagonal var-cov matrix of MM errors

Let us introduce $\bar{X}_t = (X_t', X_{t-1}')' \in \mathbb{R}^{2d}$.
Then, the transition equation (6) can be written:

$$\bar{X}_t = \Phi \bar{X}_{t-1} + \bar{\epsilon}_t, \quad \bar{\epsilon} \sim \mathrm{NID}(0, \bar{Q}) \tag{7}$$

where:

$$\Phi \equiv \begin{pmatrix} \mathbb{I} + F & -F \\ \mathbb{I} & \mathbf{0} \end{pmatrix}, \quad \bar{Q} \equiv \begin{pmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{8}$$

Therefore, introducing $M = (\mathbb{I}, \mathbf{0}) \in \mathbb{R}^{d \times 2d}$, we have the state-space model:

$$Y_t = M \bar{X}_t + \eta_t, \quad \eta_t \sim \mathrm{N}(0, H) \tag{9}$$
$$\bar{X}_t = \Phi \bar{X}_{t-1} + \bar{\epsilon}_t, \quad \epsilon_t \sim \mathrm{N}(0, \bar{Q}) \tag{10}$$

State-space representation: Advantages

This linear Gaussian state-space representation has 3 key advantages:

1. MM noise can be treated as measurement errors on the latent prices
2. Asynchronicity can be treated as a missing values problem (Corsi Peluso Audrino 2014)

3. Likelihood can be written in closed form with Kalman filter

The EM algorithm

Let

- $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ the set of observed components
- $\mathcal{X}_n = \{X_0 \ldots, X_n\}$ the set of unobserved components
- θ the set of parameters to be estimated.

Tipically, $\mathcal{L}(\theta|\mathcal{Y}_n)$ is difficult to maximize directly $\rightarrow$ EM algorithm.

EM intuition: at iteration r, EM algorithm alternates between 2 steps:

1. **E-step:** estimate $p(\hat{\mathcal{X}}_n^r)$ given $\mathcal{Y}_n, \hat{\theta}^{r-1} \rightarrow$ Kalman Filter
2. **M-step:** update $\hat{\theta}^r$ by max the expected log likelihood of the joint data

$$\underset{\theta}{\operatorname{argmax}} \, \mathrm{E}[\log \mathcal{L}(\theta|\mathcal{Y}_n, \hat{\mathcal{X}}_n^r)]$$

Useful EM properties:

- Under some regularity assumptions (Dempster et al. 1977) always increases the likelihood
- No need to compute the inverse of the Hessian as in the Newton-Raphson

The EM algorithm, cont'd

E-step: Compute the expectation of the complete log-likelihood function

$$G(\theta|\mathcal{Y}_n,\mathcal{X}_n) \equiv \mathrm{E}[\log\mathcal{L}(\theta|\mathcal{Y}_n,\mathcal{X}_n)] = -\frac{n}{2}\log|Q| - \frac{1}{2}\mathrm{Tr}[M'Q^{-1}M(C - B\Phi' - \Phi B' + \Phi A\Phi')]$$

$$-\frac{n}{2}\log|H| - \frac{1}{2}\mathrm{Tr}[H^{-1}\sum_{t=1}^{n}[(Y_t - M\bar{X}_t^n)(Y_t - M\bar{X}_t^n)' + M\bar{P}_t^n M']$$

where $\bar{X}_t^n \equiv \mathrm{E}[\bar{X}_t|\mathcal{Y}_n]$, $\bar{P}_t^n \equiv \mathrm{Cov}[\bar{X}_t|\mathcal{Y}_n]$, $\bar{P}_{t,t-1}^n \equiv \mathrm{Cov}[\bar{X}_t,\bar{X}_{t-1}|\mathcal{Y}_n]$ are computed using the Kalman filter and smoothing recursions and

$$A \equiv \sum_{t=1}^{n}(\bar{P}_{t-1}^n + \bar{X}_{t-1}^n\bar{X}_{t-1}^n), B \equiv \sum_{t=1}^{n}(\bar{P}_{t,t-1}^n + \bar{X}_t^n\bar{X}_{t-1}^n), C \equiv \sum_{t=1}^{n}(\bar{P}_t^n + \bar{X}_t^n\bar{X}_t^n)$$

M-step: Solve first-order conditions

$$\nabla_F G(\theta|\mathcal{Y}_n,\mathcal{X}_n) = 0, \quad \nabla_Q G(\theta|\mathcal{Y}_n,\mathcal{X}_n) = 0, \quad \nabla_H G(\theta|\mathcal{Y}_n,\mathcal{X}_n) = 0$$

leading to:

$$\hat{F}_r = (B_{11} - B_{12} - A_{11} + A_{12})(A_{11} + A_{22} - A_{12} - A_{21})^{-1}$$

$$\hat{Q}_r = \frac{1}{n}M(C - B\hat{\Phi}_r' - \hat{\Phi}_r B' + \hat{\Phi}_r A\hat{\Phi}_r')M'$$

$$\hat{H}_r = \frac{1}{n}\sum_{t=1}^{n}[(Y_t - M\bar{X}_t^n)(Y_t - M\bar{X}_t^n)' + M\bar{P}_t^n M']$$

$\hat{\Phi}_r$ is built using $\hat{F}_r$ and A_{ij}, B_{ij}, $i = 1, 2$ denote the four $d \times d$ principal submatrices of A and B

Consistency and asymp normality of MLE of linear Gaussian state-space models are studied under very general conditions by Douc, Moulines & Stoffer (2014).

The 2 essential conditions are:

- **Stability**: stationarity of latent returns $\Rightarrow$ eigenvalues of F lie inside the unit circle
- **Identifiability**, the model is fully identified since the selection matrix M (coupling the observed vector Y_t to the latent vector $\bar{X}_t$) is fixed.

Under these conditions, denoting by θ_0 the true parameters, the MLE $\hat{\theta}_n$ is consistent and, as $n \to \infty$:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}[0, \mathcal{I}(\theta_0)^{-1}]$$

where $\mathcal{I}(\theta_0)$ is the Fisher information matrix

Estimation Summary

Kalman-EM gives us:

- $\widehat{X}_t$ the latent price with partial adjustments
- $\widehat{Q}$ the contemporaneous var-cov matrix of latent price X_t (p.d. by construction)
- $\widehat{\Psi}$ the lead-lag structure among the latent prices X_t

With these, in addition, we can also recover:

- $\widehat{\Sigma}$ the var-cov of the efficient price P_t:

$$\widehat{\Sigma} = \widehat{\Psi}^{-1}\widehat{Q}\widehat{\Psi}'^{-1}$$

- $\widehat{P}_t$ the estimated dynamics of the efficient price:

$$\widehat{P}_t = \widehat{\Psi}^{-1}(\widehat{X}_t - (\mathbb{I} - \widehat{\Psi})\widehat{X}_{t-1})$$

Simulations: Robustness to Asynchronicity

Simulation design:

- Sample 2 correlated Brownian motions over a time grid of $T = 10000$ equally spaced points for $N = 250$ sample paths

- Contemporaneous correlation is set at $\rho = 0.4$.

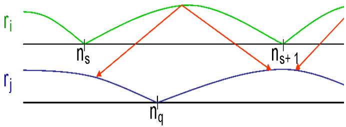
- Observations are censored using Poisson sample with missing probability $\Lambda_1 = 0.3$ and $\Lambda_2 = 0.5$

- Benchmark: the Hoffmann, Rosembaum and Yoshida (2013) estimator which applies the bivariate estimator of Hayashi-Yoshida (HY) after shifting the timestamps of one of the two series.

Hayashi and Yoshida (2005): all returns with overlap

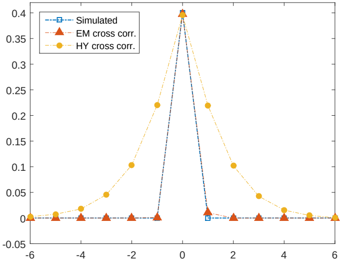
$$HY = \sum_{s=1}^{M_i} \sum_{q=1}^{M_j} r_{i,s} \; r_{j,q} \; I(\lambda_{q,s} > 0)$$

$$\lambda_{q,s} = \max(0, \min(n_{i,s+1}, n_{j,q+1} - \max(n_{i,s}, n_{j,q}))$$



- Repeat the same experiment but shifting by 1 second one of the two series

Simulations: Robustness to Asynchronicity, cont'd



Avarage correlogram estimation:

- with only contemporaneous correlation $\rho = 0.4$ (left panel)
- with shifted correlation at 1 second (right panel)

Simulations: Robustness to stochastic volatility

Simulation design:

- Misspecify constant Σ by taking $\Sigma_t = D_t R D_t$ where R a constant correlation matrix with $R_{12} = R_{21} = 0.4$ and D_t a diagonal matrix of std-dev sampled from a CIR process:

$$d\sigma_{i,t}^2 = k(\Theta_i - \sigma_{i,t}^2)dt + s\sigma_{i,t}\xi_{i,t}dt, \quad \xi_{i,t} \sim \mathrm{N}(0,1), \quad i = 1,2$$

  with $k = 10$, $s = 0.5$, $\mu \equiv \mathrm{Corr}[\xi_t, \eta_t] = 0.2$.

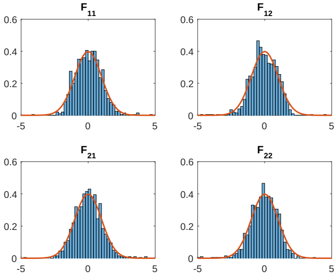
- 'lead-lag matrix':

$$F = \begin{pmatrix} 0.1 & 0.5 \\ 0.3 & 0.1 \end{pmatrix}$$

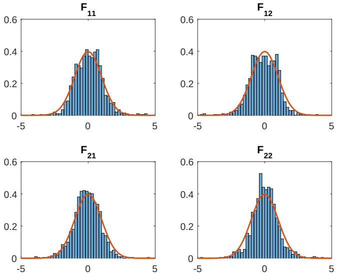
- Average signal-to-noise ratio $\bar{\delta}_i = 1$, (i.e. $h_{i,i} = q_{i,i}$)

- Poisson censoring with missing probability Λ_i, $i = 1, 2$.

Simulations: Robustness to stochastic volatility, cont'd



(a) $\delta_1 = \delta_2 = 1,\ \Lambda_1 = \Lambda_2 = 0$



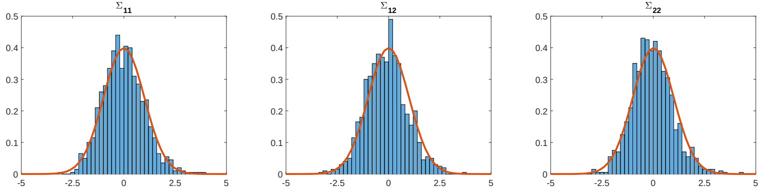
(b) $\delta_1 = \delta_2 = 1,\ \Lambda_1 = \Lambda_2 = 0.5$

Histogram of the pivotal statistic $(\hat{F}_{ij} - F_{ij})/\hat{\sigma}^F_{ij}$

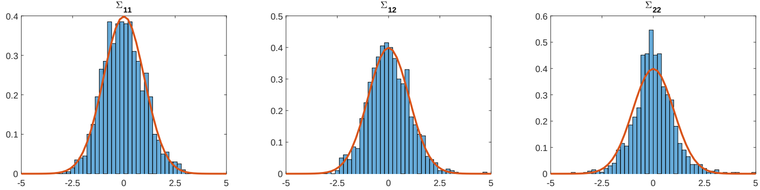
Standard normal distribution superimposed (red line).

Simulations: Robustness to stochastic volatility, cont'd

No Missings ($\bar{\delta} = 1, \Lambda = 0$)



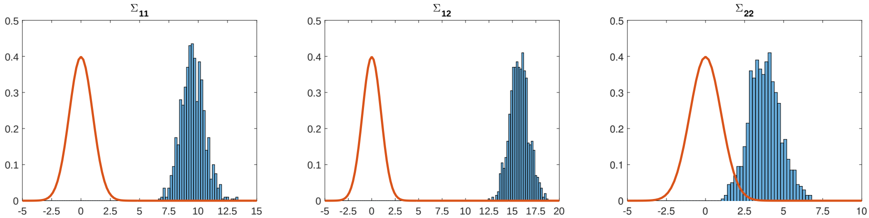
Missings ($\bar{\delta} = 1, \Lambda = 0.5$)



Histogram of the pivotal statistic $(\hat{\Sigma}_{ij} - \frac{1}{T} QV)/\hat{\sigma}^{\Sigma}_{ij}$
QV is the Quadratic Variation of the efficient price P_t.
Standard normal distribution superimposed (red line).

Simulations: HY with contemporaneous and lead-lag dependence

No Missings ($\bar{\delta} = 1, \Lambda = 0$)



Histogram of the pivotal statistic $(\hat{\Sigma}_{ij} - \frac{1}{T} QV)/\hat{\sigma}^{\Sigma}_{ij}$

for HY estimator in presence of contemporaneous and lead-lag effects

Empirical application: Data

- We consider transaction data of $d = 11$ assets traded in the NYSE in 2014: automobile (GM), banking sector (C,JPM,BAC,MS,GS) and energy sector (XOM,CVX,SLB,GM,COP,GE)
- The model is estimated on each business day of 2014. The reported correlograms are average over the whole sample.

Table: Summary statistics

Symbol	$1-\overline{\Lambda}$	$\overline{n}$	$\overline{\delta}$	Symbol	$1-\overline{\Lambda}$	$\overline{n}$	$\overline{\delta}$
XOM	0.184	4304	1.178	BAC	0.131	3079	0.328
C	0.163	3832	1.246	COP	0.120	2828	0.494
JPM	0.160	3743	0.999	GE	0.108	2543	0.641
CVX	0.152	3553	0.850	MS	0.103	2416	0.741
SLB	0.147	3454	0.613	GS	0.080	1873	0.630
GM	0.134	3135	0.888	-	-	-	-

Empirical application: Banking sector

	Group I				
	avg F_{ij}				
	C	JPM	BAC	MS	GS
C	0.0886****	0.0472****	0.0220*	-0.0635****	0.1276****
JPM	0.0318***	0.1023****	0.0065$^{(ns)}$	-0.0709****	0.1358****
BAC	0.0518****	0.0657****	0.0863****	0.0201$^{(ns)}$	0.1040****
MS	0.0752****	0.0973****	0.0107$^{(ns)}$	0.0193*	0.1542****
GS	0.0334**	0.0011$^{(ns)}$	-0.0031$^{(ns)}$	-0.0743****	0.1647****

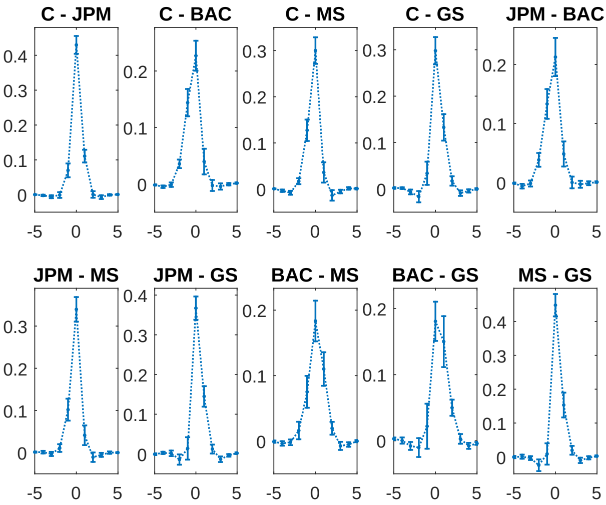
p-value of the t-test: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$, $^{****}p \leq 0.0001$, $^{(ns)}p > 0.05$.

Empirical application: Energy sector + GM

	Group II					
	avg F_{ij}					
	XOM	CVX	SLB	GM	COP	GE
XOM	0.0776^{****}	0.0428^{***}	0.0273^{****}	-0.0143^{**}	$-0.0263^{(ns)}$	-0.0408^{****}
CVX	0.0232^{*}	0.0981^{****}	0.0155^{*}	$-0.0137^{(ns)}$	-0.0327^{*}	-0.0322^{***}
SLB	$0.0135^{(ns)}$	0.0665^{***}	0.0946^{****}	-0.0251^{*}	-0.0709^{****}	-0.0355^{*}
GM	0.0809^{****}	0.0657^{**}	0.0733^{****}	0.0686^{****}	$-0.0035^{(ns)}$	$0.0182^{(ns)}$
COP	0.0318^{**}	0.0502^{***}	0.0485^{****}	$-0.0114^{(ns)}$	0.0531^{****}	$-0.0045^{(ns)}$
GE	0.0321^{*}	0.0542^{***}	0.0424^{****}	$0.0137^{(ns)}$	$-0.0015^{(ns)}$	0.0585^{****}

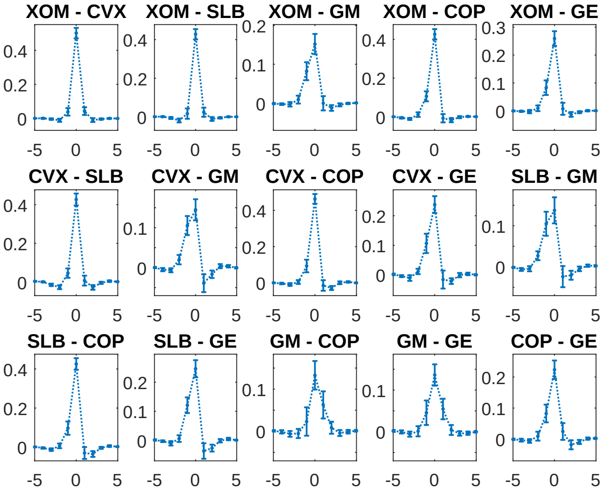
p-value of the t-test: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$, $^{****}p \leq 0.0001$, $^{(ns)}p > 0.05$.

Empirical application: Banking sector



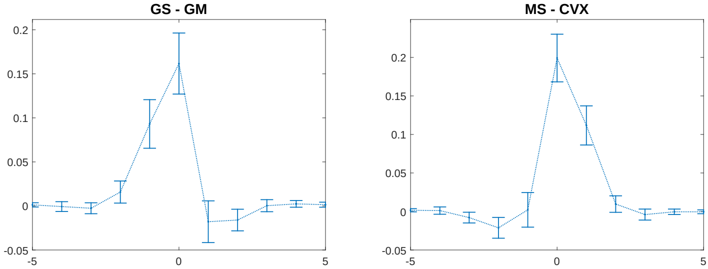
Lead-lag correlations among assets belonging to the banking sector

Empirical application: Energy sector + GM



Lead-lag correlations among assets belonging to the energy sector

Empirical application: Across sectors



Cross autocorrelations between stocks belonging to different sector

Conclusions

- Describe HF lead-lag effects within a theoretical framework which extends well established microstructure model of partial price adjustments.
- Propose an estimation procedure for lead-lag correlation in the latent price process robust to:
  - Asynchronous trading
  - Market microstructure noise

  $\Rightarrow$ which can be seen as a Granger test on latent variables
- Provide estimator for the Integrated Covariance of efficient price robust to:
  - Market microstructure noise
  - Asynchronous trading
  - Lead-lag dependence
  - And p.d. by construction
- Empirical application finds significant lead-lag effects in equity data supporting the hypothesis of multivariate price formation process.