Covariance Measurement in the Presence of Non-Synchronous Trading and Market Microstructure Noise

Jim E. Griffin and Roel C.A. Oomen[*]

June 27, 2006

Abstract

This paper studies the problem of covariance estimation when price observations are subject to non-synchronicity and contaminated by i.i.d. microstructure noise. We derive closed form expressions for the bias and variance of three popular covariance estimators, namely realized covariance, realized covariance plus lead- and lag-adjustments, and the Hayashi and Yoshida estimator, and present a comprehensive investigation into their properties and relative efficiency. The key finding of this paper is that the ordering of covariance estimators in terms of efficiency depends crucially on the level of microstructure noise. In fact, for sufficiently high levels of noise, the standard realized covariance estimator (without any corrections for non-synchronous trading) can be most efficient. An empirical illustration using TAQ quote and transaction data confirms the validity of our methodology and points to some avenues for future research.

[*] Griffin is from the Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom. E-mail: j.e.griffin@warwick.ac.uk. Oomen is from the Department of Finance, Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom. E-mail: roel.oomen@wbs.ac.uk. Oomen is also a research affiliate of the Department of Quantitative Economics at the University of Amsterdam, The Netherlands. The authors wish to thank Neil Shephard, the seminar participants at Merrill Lynch New York, the University of Maastricht, and the 2006 CIREQ “realized volatility” conference in Montreal for helpful comments and suggestions.

1 Introduction

The covariance structure of asset returns is fundamental to many issues in finance, and the importance of accurate covariance estimation can therefore hardly be understated. In recent years, high frequency data have become increasingly available for a wide range of securities (including all publicly traded US stocks, numerous exchange rates, treasury bonds, etc) and together with this, we have witnessed a shift in focus away from parametric conditional covariance estimation based on daily or weekly data to the model-free ex-post measurement of realized quantities based on intra-day data (e.g. Andersen and Bollerslev, 1998; Andersen, Bollerslev, Diebold, and Labys, 2003; Barndorff-Nielsen and Shephard, 2004). While in theory the efficiency gains associated with high frequency data are often considerable, particularly for variance/covariance estimation, practical implementation is faced with the complications that arise from the emerging market microstructure noise effects which contaminate observed prices when sampled at high frequency. Early recognition of this issue is provided by Niederhoffer and Osborne (1966) who document substantial serial correlation in returns that can, to a large extent, be attributed to the presence of a bid-ask spread. More recently, the impact of microstructure noise has been studied extensively in the context of realized variance measurement, see for instance Aït-Sahalia, Mykland, and Zhang (2005); Bandi and Russell (2006); Barndorff-Nielsen, Hansen, Lunde, and Shephard (2006); Corsi, Zumbach, Müller, and Dacorogna (2001); Hansen and Lunde (2006); Oomen (2006); Zhang, Mykland, and Aït-Sahalia (2005). The main finding of this literature is that microstructure noise makes realized variance a biased and inconsistent estimator for the integrated variance and various approaches have been suggested to deal with this, including sparse sampling, pre-filtering, bias correction, time deformation, and alternative “second generation” realized variance measures based on kernel smoothing or subsampling.

When turning to the multivariate problem of realized covariance measurement, matters don’t simplify because, in addition to noise, the impact of non-synchronous trading becomes a real concern. In short, when the arrival times of trades are random and hence non-synchronous across assets, returns sampled at regular intervals in calendar time will correlate with preceding and successive returns on other assets, even when the underlying correlation structure is purely contemporaneous. This is known as the Fisher effect (Fisher, 1966). Moreover, when the sampling interval is reduced, the covariance between any two asset return series matched in calendar time diminishes and, in the limit, converges to zero. This observation has first been made by Epps (1979). Motivated by these profound consequences of non-synchronous trading[1], a number of alternative covariance estimators have been suggested in the literature. Scholes and Williams (1977) modify the standard covariance estimator by adding the first lead and lag of the sample autocovariance. Dimson (1979) and Cohen, Hawawini, Maier, Schwartz, and Whitcomb (1983) generalize this estimator to include k leads and lags. Here the choice of how many leads and lags to include is determined by trading off a bias reduction against an increase in the variance of the estimator when increasing k. More recently, Hayashi and Yoshida (2005) propose a covariance estimator that is computed by accumulating the cross-product of all fully and partially overlapping transaction returns (see also de Jong and Nijman, 1997, for a similar estimator). Importantly, this estimator is free of any biases due to non-synchronous trading but it does require the exact timing of transactions. Clearly, if these are not available, one may still be forced to rely on the conventional realized covariance measure, with or without lead-lag adjustments, as a necessary compromise.

The contribution this paper makes is to join the above two streams of literature, and analyze the properties of the realized covariance (RC), the realized covariance plus lead-lag adjustment (RCLL), and the Hayashi-Yoshida covariance estimator (HY) in a setting with non-synchronous trading *and* market microstructure noise contaminations. While both issues have been analyzed extensively in isolation, with the literature on variance estimation focusing on microstructure noise and the literature on covariance estimation focussing on the non-synchronicity of trades, the combined impact of both these effects is clearly of interest. Besides, there is no reason to expect that the impact of noise on covariance estimation will be the same as that on variance estimation, and a separate investigation of this issue is thus warranted. With regard to microstructure noise, we employ an i.i.d. specification that is standard in the RV literature (see for instance Bandi and Russell, 2006; Zhang, Mykland, and Aït-Sahalia, 2005). To generate non-synchronicity of trades, we assume that the transaction times are generated by independent Poisson processes, with an arrival intensity that can vary across assets but is constant over time (see e.g. Hayashi and Yoshida, 2005). In this setting, we present closed form expressions for the bias and variance of the RC, RCLL, and HY covariance estimators. We provide a detailed discussion of the relative efficiency

[1] The impact of non-synchronous trading has also received considerable attention in the literature on beta estimation for asset pricing (e.g. Shanken, 1987), index autocorrelation (e.g. Atchison, Butler, and Simonds, 1987; Lo and MacKinlay, 1990), and lead-lag patterns (Chordia and Swaminathan, 2000; de Jong and Nijman, 1997).

of these estimators and also discuss optimal sampling. Our findings can be summarized as follows: (i) i.i.d. noise does not bias any of the covariance estimators but it does make them inconsistent, (ii) as predicted by the Epps effect, non-synchronous trading leads to a substantial downward bias in realized covariance, which can be extenuated by inclusion of lead-lag adjustments, (iii) a careful choice of sampling frequency is crucial for all covariance estimators including HY (with i.i.d. microstructure noise the variance of the HY estimator increases so that for sufficiently high levels of noise HY is best implemented with aggregated transaction returns) and (iv) the ordering of competing covariance estimators in terms of their efficiency crucially depends on the level of noise. This last point is perhaps the most surprising finding because it suggests that the "plain-vanilla" RC estimator can – in certain circumstances – attain greater efficiency than either RC with lead-lag adjustments or the HY estimator. We also present some empirical results, and find that they confirm the validity of our theoretical framework. In particular, we find that the rate of decay in the empirical covariance signature plots for RC is broadly consistent with that implied by our theory. For the HY estimator we detect a systematic downward bias which, we conjecture, is caused by sluggish adjustment of prices.

To conclude, it is emphasized that a number of recent papers have addressed similar issues to those studied here. For instance, Martens (2004) uses simulations to investigate the relative performance of alternative covariance estimators, including RC, RCLL, and HY, in the context of the Lo and MacKinlay (1990) non-synchronous trading model. Bandi and Russell (2005) provide a formal analysis of realized covariance in the presence of noise (but abstract from the non-trading issue) whereas Sheppard (2005) introduces the concept of "scrambling" to study non-synchronicity and realized covariance (but abstracts from the noise issue). Zhang (2006) provides an analytic treatment of the RC estimator in a general framework that includes non-synchronous trading and microstructure noise, Hayashi and Yoshida (2006) study the joint distribution of HY covariance estimator and RV in the absence of noise, Corsi (2006) studies the HY estimator using the HAR model allowing for time varying covariance structure, while Voev and Lunde (2005) use simulations to study the properties of the RC and HY estimators (and extensions thereof) for general noise dependence and non-trading scenarios. This paper is distinguished from the above literature in that it presents a comprehensive investigation of the three competing covariance estimators simultaneously within a unified framework that incorporates both non-synchronous trading and microstructure noise. Because closed form expressions for the bias and MSE are available, the relative efficiency of the RC, RCLL and HY estimators can be studied.

The remainder of this paper is organized as follows. Section 2 introduces the modeling framework that incorporates non-synchronous trading and microstructure noise. This is then used to study the properties of RC in section 2.1, RCLL in section 2.2, and HY in section 2.3. A detailed investigation into the relative efficiency of the covariance estimators can be found in section 2.4. Section 3 presents an empirical illustration using NYSE quote and transaction data for five randomly selected Dow Jones 30 components. It also points at some directions for future research. Section 4 concludes and proofs are collected in the Appendix.

2 Covariance estimation with non-synchronous and noisy returns

Let $S^{(j)}(t)$ denote the time$-t$ efficient (logarithmic) price of asset j, for $t \in [0,1]$. It is assumed that prices of asset j are observed at a set of discrete times $\{t_m^{(j)}\}_{m=1}^{M_j}$ with $0 \leq t_1^{(j)} < \ldots < t_{M_j}^{(j)} \leq 1$ and are subject to observation error:

$$p_m^{(j)} = s_m^{(j)} + u_m^{(j)} \qquad \text{for} \quad m = 1, \ldots, M_j \tag{1}$$

where $s_m^{(j)} = S^{(j)}(t_m^{(j)})$ and $u_m^{(j)}$ is a "noise" process to be specified. In practice, the observation times typically correspond to the occurrence of transactions or quote-revisions whereas the observation noise is due to market microstructure effects such as the bid-ask spread. Thus, the efficient price process is latent and all inference about the process in general, and the variance/covariance structure in particular, is necessarily based on the discretely sampled and noisy observations p. Throughout the remainder of this paper we make the following assumptions:

Assumption 1 [Brownian motion] *The efficient price process S is a correlated Brownian motion, i.e. $S^{(j)} = \sigma_j W^{(j)}$ with $dW^{(i)}dW^{(j)} = \rho_{ij}dt$.*

Assumption 2 [Poisson sampling] *The observation times of asset j, i.e. $\{t_m^{(j)}\}_{m=1}^{M_j}$, are generated by a Poisson process with intensity λ_j, and are independent of observation times of other assets.*

Assumption 3 [I.I.D. noise] *The noise process $u^{(j)}$ is i.i.d. $(0, \xi_j^2)$ and independent of the efficient price process.*

In the above, the efficient price is specified as a martingale allowing for contemporaneous correlation between the different assets. The two salient features that are central to this paper, namely non-synchronicity of observation times (also referred to as non-synchronous trading or non-trading) and microstructure effects are captured through the independent Poisson sampling and i.i.d. noise specification respectively. Note that Assumptions 1 and 2 constitute a special case considered by Hayashi and Yoshida (2005) whereas assumption 3 is standard in the realized variance literature (see for instance Bandi and Russell, 2006; Zhang, Mykland, and Aït-Sahalia, 2005). Of course, the specification of the price, noise, and sampling processes necessarily reflects a balance between generality and analytic tractability and constitutes, at best, a first order approximation of reality. Still, it should be pointed out that the assumptions may not be as restrictive as they appear at first sight for at least two reasons, namely (i) seemingly dependent noise may often arise as an artefact of the sampling scheme, even when the actual noise process is i.i.d. (see Griffin and Oomen, 2005, for further discussion) and (ii) non-homogeneity of trade arrivals and stochastic volatility can be accounted for by appropriately deforming the time scale. The implicit independence between the price innovations and the trade arrival process is the more restrictive assumption but, as discussed by de Jong and Nijman (1997), is difficult to relax in the current context.

2.1 Realized covariance

To compute realized covariance (RC), the multivariate price process needs to be sampled on a common grid. In this paper we assume that the "previous tick" method is used where at each sampling point the most recently observed price for each asset is recorded, i.e.

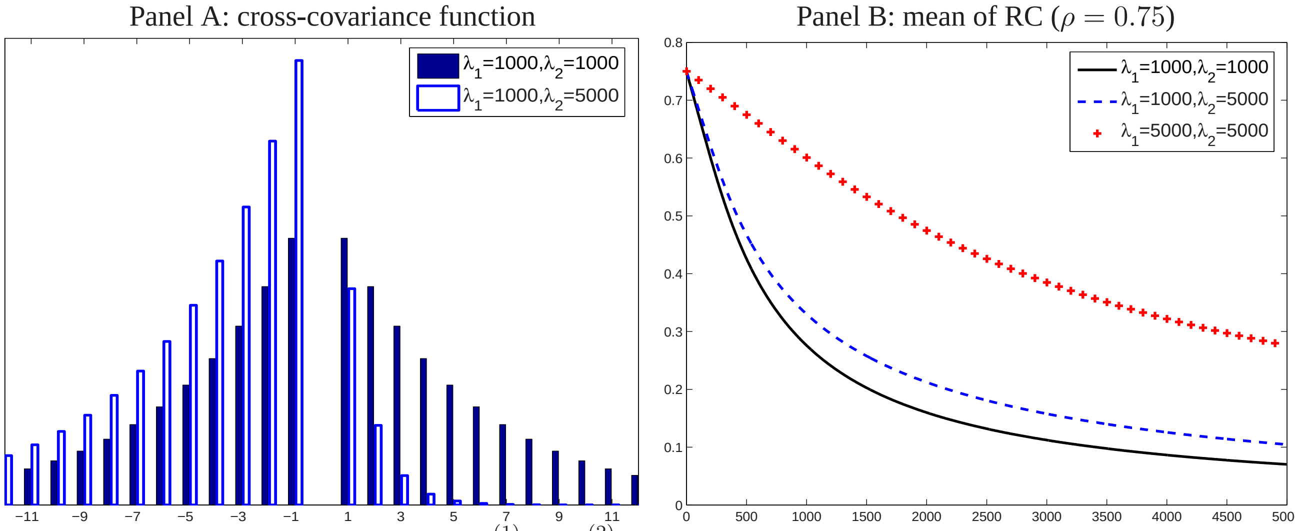
$$P_t^{(j)} = p_{N_j(t)}^{(j)} \qquad \text{where } N_j(t) = \sup_n \{n | t_n^{(j)} \leq t\}$$

It is important to emphasize that sampling prices in this fashion does *not* eliminate the non-trading problem but merely ensures that returns across assets are measured over matching intervals. For ease of exposition, we focus on two assets only, i.e. $j \in \{1, 2\}$ and use ρ as a shorthand for ρ_{ij}. The object of econometric interest in the covariance estimation is thus $\rho\sigma_1\sigma_2$. With M returns sampled at regular intervals $\Delta = 1/M$, the RC for asset 1 and 2 is computed as:

$$RC_M = \sum_{m=1}^{M} r_m^{(1)} r_m^{(2)} \tag{2}$$

where $r_m^{(j)} = P_{m/M}^{(j)} - P_{(m-1)/M}^{(j)}$ (suppressing dependence on M). In the absence of noise and non-synchronous trading, we have that $E(RC_M) = \rho\sigma_1\sigma_2$ and $V(RC_M) = M^{-1}(1 + \rho^2)\sigma_1^2\sigma_2^2$ so that RC is unbiased and

Figure 1: Impact of non-synchronous trading on cross correlations and RC



Panel A: cross-covariance function

Panel B: mean of RC ($\rho = 0.75$)

Note. Panel A plots the correlation between $r_m^{(1)}$ and $r_{m+h}^{(2)}$ (i.e. Eq. 4) as a function of h. Panel B plots the mean of RC, given by Eq. (5) as a function of M.

consistent when $M \rightarrow \infty$. See Barndorff-Nielsen and Shephard (2004) for a comprehensive treatment of the asymptotic distribution theory of RC for continuous semi-martingales. When adding i.i.d. noise to the process through assumption 3, RC remains unbiased (unlike realized variance!) but is now inconsistent:

$$V(RC_M) = M^{-1}(1+\rho^2)\sigma_1^2\sigma_2^2 + 2\sigma_1^2\xi_2^2 + 2\sigma_2^2\xi_1^2 + (6M-2)\,\xi_1^2\xi_2^2. \tag{3}$$

Note that when only one asset is contaminated with noise, RC remains inconsistent but it is optimal to sample as frequent as possible from a MSE criterion viewpoint. See Bandi and Russell (2005) and Voev and Lunde (2005) for further discussion of the impact of noise on RC.

When introducing non-synchronicity of price observations through assumption 2, regularly sampled returns are no longer correlated only contemporaneously, but will also correlate with leads and lags of sampled returns of other assets. Under assumptions 1 and 2, the autocovariance function of returns can be expressed as:

$$E(r_m^{(1)} r_{m+h}^{(2)}) = \begin{cases} \rho\sigma_1\sigma_2 \frac{\lambda_1(1-e^{-\lambda_2\Delta})^2}{\lambda_2(\lambda_1+\lambda_2)} e^{-\lambda_2(h-1)\Delta} & \text{for } h > 0 \\ \rho\sigma_1\sigma_2 \frac{\lambda_2(1-e^{-\lambda_1\Delta})^2}{\lambda_1(\lambda_1+\lambda_2)} e^{-\lambda_1(h-1)\Delta} & \text{for } h < 0 \end{cases} \tag{4}$$

See Appendix A.2 for details on the derivation. To illustrate the lead-lag dependence due to non-trading, we plot the covariance function of returns in Eq. (4) with sampling frequency $M = 5000$ and for two scenarios of the observation arrival frequency, namely (i) $\lambda_1 = \lambda_2 = 1000$ and (ii) $\lambda_1 = 1000, \lambda_2 = 5000$. The result can be found in Panel A of Figure 1. Because the number of sampled returns M is large relative to the number of observations, cross correlations are substantial and extend to a large number of leads and lags. As expected, with equal arrival rates the dependence structure is symmetric. When increasing the arrival rate on asset 2, the lead dependence is reduced relative to the lag dependence: the non-synchronicity is now primarily caused by asset 1, hence the asymmetric pattern. The following result makes the impact of non-synchronous trading on RC explicit.

Theorem 2.1 *Given Assumptions 1, 2, and 3, the expectation of RC is equal to:*

$$E(RC_M) = \rho\sigma_1\sigma_2\beta_M, \tag{5}$$

where

$$\beta_M = 1 - \frac{M}{\lambda_1+\lambda_2}\left(\frac{\lambda_1}{\lambda_2}\mu_2 + \frac{\lambda_2}{\lambda_1}\mu_1\right),$$

and $\mu_j = 1 - e^{-\lambda_j/M}$. *The variance of RC is equal to:*

$$\begin{aligned} V\left(RC_M\right) \quad = \quad & \frac{\left(1+2\rho^2\right)\sigma_1^2\sigma_2^2}{M} + 4M\frac{\rho^2\sigma_1^2\sigma_2^2}{\lambda_1+\lambda_2}\left(\frac{\mu_1\mu_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_2^2}\mu_2 + \frac{\lambda_2}{\lambda_1^2}\mu_1 - \frac{1}{M}\left(\frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1}\right)\right) \\ & - \frac{\rho^2\sigma_1^2\sigma_2^2\beta_M^2}{M} + 2\mu_2\sigma_1^2\xi_2^2 + 2\mu_1\sigma_2^2\xi_1^2 + 4M\mu_1\mu_2\xi_1^2\xi_2^2 + 2\left(M-1\right)\mu_1^2\mu_2^2\xi_1^2\xi_2^2. \end{aligned} \tag{6}$$

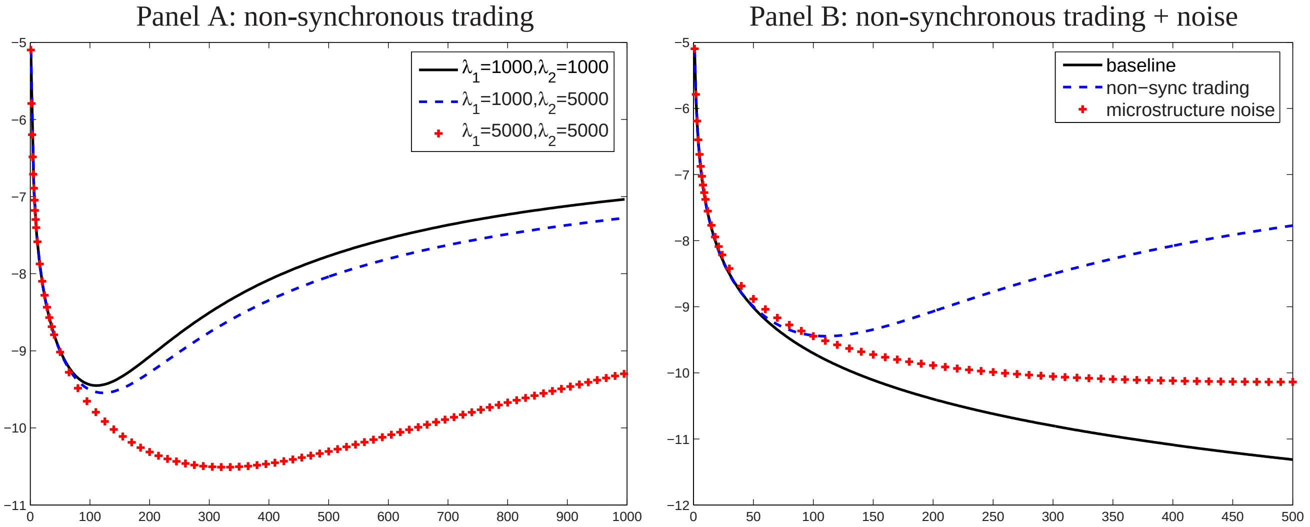
Proof See Appendix A.1. ■

Eq. (5) indicates that non-synchronous trading makes RC a biased estimator for the covariance and the Epps effect is apparent: because $\lim_{M\to\infty}\beta_M = 0$, the expectation of RC tends to zero when the sampling frequency increases. In fact, since[2] $\lim_{M\to\infty} V(RC) = 0$, RC is equal to zero with certainty in the limit, even in the presence of noise! To illustrate the magnitude of the bias, Panel B of Figure 1 plots the expectation of RC as a function of the number of sampled returns M, for various combinations of λ. As expected, when λ is small relative to M, the impact of non-synchronicity is more prominent than when λ is large relative to M. Also, the magnitude of the bias is primarily determined by the slowest trading asset, which in itself has of course important implications for high dimensional covariance measurement in practice.

Next, using the variance expression in Eq. (6), we draw the mean squared error (MSE) of RC as a function of M in Figure 2. From Panel A we can see that the minimum attainable MSE is largely determined by the

[2] When $M\to\infty$, then $\beta_M \to 0, \mu_j \to 0, M\mu_j \to \lambda_j$

Figure 2: Mean squared error of RC



Panel A: non-synchronous trading

Panel B: non-synchronous trading + noise

Note. Panels A and B draws the log MSE of RC as a function of M using the mean and variance expressions in Theorem 2.1.

slowest trading asset. In particular, the optimal sampling frequency (i.e. the choice of M that minimizes the MSE criterion), remains roughly unchanged when λ_2 is increased and keeping λ_1 fixed. Of course, increasing the trade intensity of both assets leads to a higher optimal sampling frequency and a lower minimum MSE. Because of the rather lengthy and complicated variance expression it is not possible to characterize the optimal sampling frequency in closed form, but in practice it can be obtained numerically in a straightforward fashion.

To isolate the impact of non-synchronous trading and microstructure noise, Panel B draws the MSE of RC under the following three scenarios (i) absence of non-synchronicity and noise, (ii) non-synchronicity only, and (iii) microstructure noise only (here $\xi_j = \sigma_j/\sqrt{1000}$ which implies a noise ratio of 1 when $M = 1000$ and is in line with empirical estimates for transaction data of US large cap stocks, see Oomen, 2006). As expected, under scenario (i) the MSE decays monotonically in M because RC is unbiased and consistent. Comparing the MSE under scenarios (ii) and (iii) indicates that the impact of noise is quite different from that of non-synchronous trading. Specifically, while the penalty of sampling at (too) low frequency in terms of MSE is comparable in the presence of noise or non-synchronous trading, the penalty of sampling at (too) high frequency is much more severe for non-synchronous trading. Put differently, a careful choice of sampling frequency is paramount with non-trading.

2.2 Realized covariance plus leads and lags

A natural way to mitigate the biases induced by non-synchronicity is to add leads and lags of the empirical autocovariance function of returns to the realized covariance measure. This approach was first proposed by Scholes and Williams (1977), and later extended by Dimson (1979), and Cohen, Hawawini, Maier, Schwartz, and Whitcomb (1983). The modified covariance estimator with lead lag adjustment (RCLL) is specified as:

$$RCLL_M = \sum_{m=1}^{M} \sum_{l=-L}^{U} r_{m+l}^{(1)} r_{m}^{(2)}. \tag{7}$$

Considering Panel A of Figure 1, it is quite intuitive that such a lead-lag correction can be effective in reducing the bias of RC. In the absence of non-synchronous trading, the optimal choice for U and L is zero as the leads and lags don't contain any useful information regarding the contemporaneous covariance structure. When there is non-synchronicity, however, the leads and lags are informative and the choice of U and L will be determined by trading off a bias reduction against an increase in variance of the estimator. Recent empirical application of the RCLL estimator can be found in Bollerslev and Zhang (2003) and Bandi and Russell (2005).

Theorem 2.2 *Given Assumptions 1, 2, and 3, the expectation of RCLL is equal to:*

$$E(RCLL_M) = \rho\sigma_1\sigma_2\beta_M^*, \tag{8}$$

where

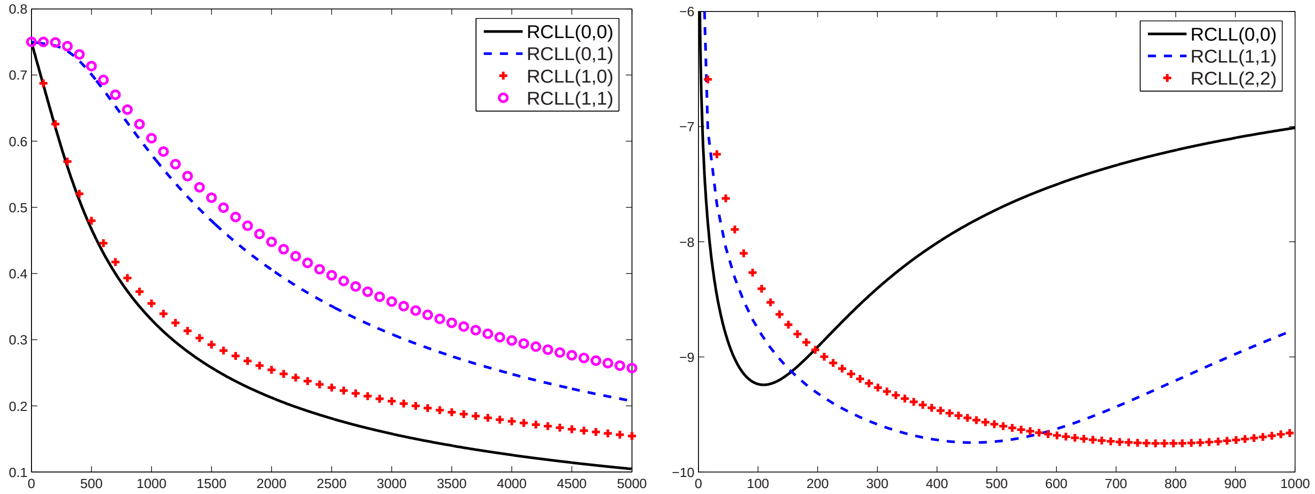
$$\beta_M^* = 1 - \frac{M}{\lambda_1+\lambda_2}\left(\frac{\lambda_1}{\lambda_2}\mu_2 e^{-\lambda_2 L\Delta} + \frac{\lambda_2}{\lambda_1}\mu_1 e^{-\lambda_1 U\Delta}\right).$$

The variance of RCLL is equal to:

$$\begin{aligned} V\left(RCLL_M\right) \;=\; & \frac{\sigma_1^2\sigma_2^2(U+L+1)}{M} + 2M\rho^2\sigma_1^2\sigma_2^2\mu_1^*\mu_2(p_1\alpha_1+p_2\alpha_2+p_3\alpha_3) - \frac{\rho^2\sigma_1^2\sigma_2^2\beta_M^{*2}}{M} \\ & +2\left(M-1\right)\left(\mu_1^*\mu_1+\mu_1^*-\mu_1\right)\mu_2^2\xi_1^2\xi_2^2 - 2\left(M-1\right)\mu_2^2\xi_2^2\sigma_1^2\left(U+L\right)\Delta \\ & +2\mu_2\sigma_1^2\xi_2^2\left(U+L+1\right) + 2\mu_1^*\sigma_2^2\xi_1^2 + 4M\mu_1^*\mu_2\xi_1^2\xi_2^2 \end{aligned} \tag{9}$$

where $\mu_1^* = 1 - e^{-\lambda_1\Delta(U+L+1)}$, $p_1 = (1-e^{-U\lambda_1\Delta})/\mu_1^*$, $p_2 = \mu_1 e^{-U\lambda_1\Delta}/\mu_1^*$, $p_3 = \frac{\lambda_1}{\lambda_1+\lambda_2}(1-e^{-L(\lambda_1+\lambda_2)\Delta})$

Figure 3: Bias and MSE of RCLL


Panel A: Bias

Panel B: MSE

Note. Panel A (B) plots the expectation (log MSE) of RCLL(L,U) as a function of M for various choices of U and L. In Panel A we set $\lambda_1 = 1000, \lambda_2 = 5000$ and $\rho = 0.75$. In Panel B we set $\lambda_1 = \lambda_2 = 1000$, $\rho = 0.75$, $\sigma_1 = \sigma_2$, and $\xi_1 = \xi_2 = \sigma_1/\sqrt{\lambda_1}$.

$e^{-(U+1)\lambda_1\Delta}/\mu_1^*$, *and*

$$\begin{aligned}
\alpha_1 &= \frac{2}{\lambda_2^2} + \frac{\Delta^2}{\mu_2} - 2\lambda_1 \frac{\Delta\left(L\mu_2+1\right)\left(\lambda_1+\lambda_2\right)+\mu_2}{\lambda_2\left(\lambda_1+\lambda_2\right)^2\mu_2} e^{-\lambda_2 L\Delta} \\
\alpha_2 &= \frac{\Delta^2}{\mu_1\mu_2} + \frac{2}{\lambda_1+\lambda_2}\frac{1}{\mu_1\mu_2}\left(\frac{\mu_1\mu_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_2^2}\mu_2 + \frac{\lambda_2}{\lambda_1^2}\mu_1 - \Delta\left(\frac{\lambda_1}{\lambda_2}+\frac{\lambda_2}{\lambda_1}\right)\right) \\
&\quad +2\frac{\lambda_1\left(\lambda_1+\lambda_2\right)\left(\mu_1\mu_2+\lambda_2\Delta\right) - \lambda_1^2\mu_2 - \lambda_2^2\mu_1}{\mu_1\mu_2\left(\lambda_1+\lambda_2\right)^2\lambda_2^2}\left(1-e^{-\lambda_2 L\Delta}\right) - 2\frac{\lambda_1}{\lambda_2}\frac{Le^{-\lambda_2 L\Delta}\Delta}{\lambda_1+\lambda_2} \\
\alpha_3 &= \frac{2}{\lambda_2^2}\frac{1-e^{-\lambda_2 L\Delta}}{1-e^{-(\lambda_2+\lambda_1)L\Delta}} - 2\frac{L\Delta\lambda_1\left(\lambda_1+\lambda_2\right) - \lambda_2\left(1-e^{-\lambda_1 L\Delta}\right)}{\lambda_1\lambda_2\left(\lambda_1+\lambda_2\right)\left(1-e^{-(\lambda_1+\lambda_2)L\Delta}\right)} e^{-\lambda_2 L\Delta}.
\end{aligned}$$

Proof See Appendix A.2. ■

The bias of RCLL is characterized by β_M^*, from which it is clear that at a given sampling frequency M the bias can be made arbitrarily small by setting U and L sufficiently high (at the cost of higher variance). Also, with equal observation frequencies $\lambda_1 = \lambda_2$, inclusion of either a lead or a lag adjustment reduces the bias by the same amount. However, with unequal observation frequencies the effectiveness of a bias correction varies, e.g. when $\lambda_2 > \lambda_1$ a lead adjustment on asset 1 is more effective than a lag adjustment. This is illustrated in Panel A of Figure 3. The intuition is simple: when $\lambda_2 > \lambda_1$ non-synchronicity is primarily caused by the relatively infrequent trading of asset 1 so that returns of asset 1 are more likely to correlate with lagged returns of asset 2 rather than with lead returns of asset 2. Hence, a lead adjustment on asset 1, or equivalent a lag adjustment on asset 2, delivers the greater bias reduction.

To gain some insights into the efficiency of RCLL, Panel B of Figure 3 draws the MSE using the expressions in Eqs. (8) and (9). Here, the benchmark is RC as represented by the solid line. It is clear from the graph that with non-synchronous trading, a first order lead-lag adjustment enables one to sample at a higher frequency and substantially reduce the minimum attainable MSE. For the chosen parameters in this example, a second order lead-lag adjustment is of little value: while the optimal sampling frequency increases further, the reduction in minimum attainable MSE is negligible. Of course, when U and L are set sufficiently high it can happen that RC attains a lower MSE than RCLL because the bias reduction due to the lead-lag adjustment is more than offset by an increase in the variance of the estimator (see Section 2.4 for further comparisons between the performance of RC and RCLL). In practice, the optimal sampling frequency M and optimal choice of lead-lag adjustment U and L can be determined in a straightforward fashion using Eqs. (8) and (9), albeit that they need to be determined numerically since closed form expressions are not available.

To conclude, we point out that the RCLL estimator in Eq. (7) is similar in spirit to the bias-corrected realized variance measure of Zhou (1996) or the kernel based realized variance measure of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2006). Thus, a natural extension of the estimator would take the form:

$$RCLL_M^{\omega} = \sum_{m=1}^{M} \sum_{l=-L}^{U} \omega(l) r_{m+l}^{(1)} r_m^{(2)}. \tag{10}$$

where $\omega(\cdot)$ specifies the kernel weights. A nice property of the above estimator is that it can be combined with a kernel-based RV estimator to deliver a positive definite covariance matrix. Also, the empirical analysis below suggests that "sluggish" price adjustment exacerbates the lead-lag dependence of returns in practice, and a kernel-based RC estimator such as the one above is sufficiently flexible to counter such effects. Within our framework it is possible to extend Theorem 2.2 to include general kernel weight, albeit with considerable complexity of notation. Moreover, the choice of optimal kernel is a non-trivial issue and this is therefore left for future research.

2.3 Hayashi-Yoshida covariance estimator

In a recent paper Hayashi and Yoshida (2005) propose[3] a new covariance estimator that is free of any non-trading bias and can be computed directly with observed prices, without first sampling them on a common grid as is required for RC and RCLL. The HY covariance estimator is specified as follows:

$$HY = \sum_{i=1}^{M_1} \sum_{j \in A_i} R_i^{(1)} R_j^{(2)} \tag{11}$$

where $R_i^{(j)} = p_i^{(j)} - p_{i-1}^{(j)}$ and $A_i = \{j | (t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)}) \neq \emptyset\}$. In words, HY accumulates the cross-product of all fully and partially overlapping returns. Here, the returns are sampled at the highest available observation frequency, and are therefore irregularly spaced in calendar time and asynchronous across assets.

Theorem 2.3 *Given Assumptions 1, 2, and 3, the expectation of HY is equal to:*

$$E(HY) = \rho\sigma_1\sigma_2. \tag{12}$$

The variance of HY is equal to:

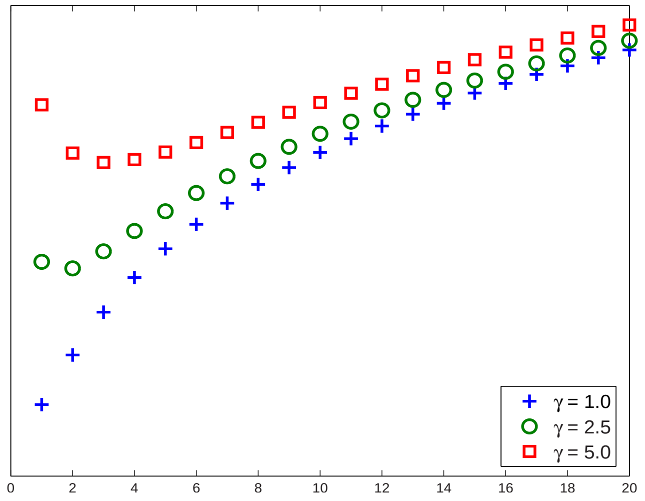
$$V(HY) = 2\sigma_1^2\sigma_2^2 \frac{\lambda_1 + \lambda_2}{\lambda_1\lambda_2} + 2\frac{\rho^2\sigma_1^2\sigma_2^2}{\lambda_1 + \lambda_2}\left(\frac{\lambda_2}{\lambda_1} + \frac{\lambda_1}{\lambda_2}\right) + 2\sigma_1^2\xi_2^2 + 2\sigma_2^2\xi_1^2 + 4\xi_1^2\xi_2^2\frac{\lambda_1\lambda_2}{\lambda_1 + \lambda_2}. \tag{13}$$

Proof See Appendix A.3. ■

In the absence of noise, unbiasedness of the HY estimator is not surprising and has already been discussed in detail by Hayashi and Yoshida (2005). Moreover, Eq. (13) suggests that the HY estimator is consistent when $\lambda_1, \lambda_2 \to \infty$, i.e. the higher the observation frequency of the process, the higher the accuracy of the HY estimator. Keep in mind, however, that in this limiting case the non-synchronicity issue disappears and the HY estimator reduces to RC. With i.i.d. noise, the HY estimator remains unbiased but is now inconsistent. Interestingly, depending on the level of noise, it may not be optimal to sample prices at the highest available observation frequency because this leads to an accumulation of noise that more than offsets the gains from using more data.

[3] The HY estimator is also studied by Hayashi and Kusuoka (2004) in a more general semi-martingale setting. Hayashi and Yoshida (2006) establish joint asymptotic normality of the HY estimator and RV. The covariance estimator of de Jong and Nijman (1997) is very similar to the one proposed by Hayashi and Yoshida, see Martens (2004) for further discussion. In independent work, Corsi (2006) proposes a "tick-by-tick realized covariance estimator" which coincides with the HY estimator. He shows that the estimator performs well, both in simulations and in practice.

Figure 4: MSE of HY estimator in presence of noise



Note. This figure plots the log MSE of the HY estimator for varying levels of microstructure noise as a function of the sampling frequency k (i.e. every k^{th} return is sampled).

To develop some further insights into this, consider the case where we sample every k^{th} observation for both assets. The variance of HY is then simply obtained by replacing λ_i with λ_i/k in the above expression, i.e.

$$V(HY_k) = 2\sigma_1^2\sigma_2^2\left(\frac{\lambda_1+\lambda_2}{\lambda_1\lambda_2} + \frac{\rho^2}{\lambda_1+\lambda_2}\left(\frac{\lambda_2}{\lambda_1}+\frac{\lambda_1}{\lambda_2}\right)\right)k + 2\sigma_1^2\xi_2^2 + 2\sigma_2^2\xi_1^2 + 4\xi_1^2\xi_2^2\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}k^{-1}.$$

From this, it follows that the optimal – MSE minimizing – aggregation or sampling frequency for the HY estimator is equal to:

$$\bar{k} = \underset{\lfloor k^*\rfloor,\lceil k^*\rceil}{\text{argmin}}\, V(HY_k), \tag{14}$$

where

$$k^* = \frac{\sqrt{2\lambda_1\lambda_2\gamma_1\gamma_2}}{\sqrt{(1+\rho^2)\left(\lambda_1^2+\lambda_2^2\right)+2\lambda_1\lambda_2}},$$

and $\gamma_i = \lambda_i\xi_i^2/\sigma_i^2$ denotes the noise ratio. The interesting case is of course when $k^* > 1$ because then it may be optimal not to use all available data but aggregate returns k^* times. This occurs when:

$$\gamma_1\gamma_2 > 1 + \frac{1}{2}\left(1+\rho^2\right)\left(\frac{\lambda_2}{\lambda_1}+\frac{\lambda_1}{\lambda_2}\right).$$

For instance, when the trade intensities and noise ratios are equal for both assets, then a sufficient condition for $k^* > 1$ is that $\gamma > \sqrt{3}$ (more generally, for $k^* > c$ we require $\gamma > c\sqrt{3}$) which is not uncommon in practice, particularly for transaction data (see e.g. Oomen, 2006). Keep in mind here that k needs to be a strictly positive integer and if we were to compute that say $k^* = 1.4$ this would not necessarily imply that we should aggregate returns because the MSE at $k = 1$ may still be smaller than at $k = 2$. Of course, when $k^* \geq 2$ aggregation is certainly optimal in this framework. To illustrate the above, Figure 4 plots the MSE of the HY estimator as a function of k for $\lambda_1 = \lambda_2 = 1000$ and varying levels of noise with γ between 1 and 5. For $\gamma = 1.0$ the MSE declines monotonically in k and the minimum MSE is attained by using data at the highest available frequency. For $\gamma = 2.5$ $(\gamma = 5.0)$ this pattern changes and the minimum MSE is attained by aggregating returns with $k = 2$ $(k = 3)$.

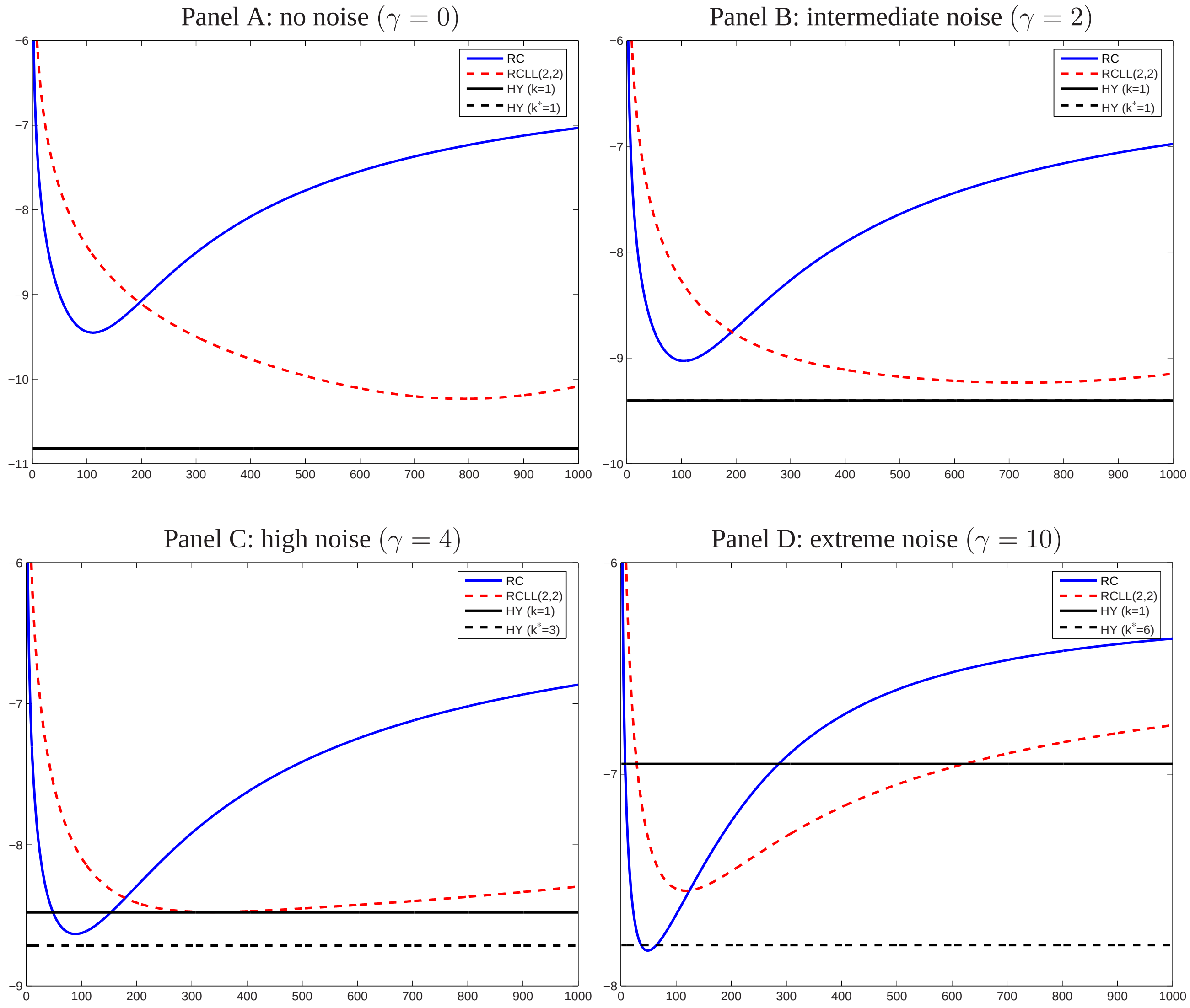
To conclude this discussion, we point out that a natural way to further improve the performance of the HY estimator is with the use of subsampling (see Zhang, Mykland, and Aït-Sahalia, 2005). One possibility would be to subsample at frequency $\bar{k}$, although the method would of course remain valid at different frequencies determined by other criteria. While an in-depth study of the properties of such a sub-sampling version of the HY estimator is of great interest, it is beyond the scope of this paper and we defer it to future research.

2.4 Relative efficiency of competing covariance estimators

The real benefit of imposing the somewhat restrictive assumptions 1, 2, and 3 above is that it allows us to derive closed form MSE expressions for all three competing covariance estimators in a unified framework. As a result, we can address the question which estimator is most efficient and under which conditions. It turns out that in this comparison the key parameter is the level of noise (i.e. γ_i) because it determines the ordering of the estimators in terms of their efficiency. The level of correlation (ρ) or asymmetries in the arrival intensity (λ_i) or level of volatility (σ_i) do not play a noteworthy role here.

Figure 5 plots the MSE of the RC, RCLL(2,2), and HY covariance estimators as a function of the sampling frequency, keeping in mind that the variance of the HY estimator is not a function of M. Motivated by the discussion above, we compute the HY estimator in the standard fashion with $k = 1$ and at its "optimal" aggregation frequency $k = k^*$. The choice of U and L for RCLL is arbitrary in this example but the results remain qualitatively the same if this is altered. First consider Panel A of Figure 5. Here, microstructure noise is absent and the relative ranking of the estimators is determined by their ability to deal with the impact of non-synchronous trad-

Figure 5: Relative efficiency of competing covariance estimators in presence of noise


Panel A: no noise $(\gamma = 0)$

Panel B: intermediate noise $(\gamma = 2)$

Panel C: high noise $(\gamma = 4)$

Panel D: extreme noise $(\gamma = 10)$

Note. This figure plots the log MSE of RC, RCLL(2,2), and HY covariance estimators as a function of sampling frequency M. The HY estimator is computed at $k = 1$ (solid horizontal line) and at $k = k^*$ (dashed horizontal line).

ing. The result is clear and unsurprising: HY performs best and RC performs worst, with RCLL improving over RC thanks to the lead-lag correction but not able to attain the efficiency of the HY estimator. In Panel B noise is introduced which leads to a narrowing of the estimators in terms of efficiency, albeit that the relative ranking remains unchanged. In Panel C the level of noise is increased, and we reached a point where RC outperforms both RCLL and the standard HY estimator! Here, the benefit of the non-trading bias correction through lead-lag adjustments (in calendar time for RCLL and in transaction time for HY) does not outweigh the associated noise accumulation, making the plain vanilla RC measure the preferred estimator. Still, if we take the level of noise explicitly into account when calculating the HY estimator, we find that $k^* = 3$ and a substantial MSE reduction can be achieved by aggregating returns. This “noise optimized” HY estimator outperforms all alternatives. Finally, in Panel D the level of noise is increased to $\gamma = 10$ in an attempt to further accentuate the behavior of the estimators. Compared to case without noise (Panel A), the situation is now completely reversed with RC outperforming RCLL and RCLL outperforming the standard HY estimator. Optimizing the HY estimator by aggregating returns to $k^* = 6$ substantially reduces the MSE but the simple RC estimator cannot be beaten.

So all in all, the main finding here is that the level of microstructure noise determines the relative efficiency of the competing covariance estimators. This situation stands in sharp contrast to that of RV where a bias correction generally improves matters (see e.g. Barndorff-Nielsen, Hansen, Lunde, and Shephard, 2006; Oomen, 2005; Zhang, Mykland, and Aït-Sahalia, 2005). For covariance estimation, the non-trading bias correction comes with a noise accumulation and it is the balancing of this trade-off that determines the relative efficiency of the estimators.

3 Empirical illustration: covariance signature plots

In order to illustrate some of the issues discussed above, and to gauge the impact of non-synchronous trading in practice, we now turn to some descriptive analysis of TAQ data for five randomly selected large cap companies. Here, the focus is on simple covariance signature plots (that is, the average covariance estimates as a function of the sampling frequency) because this will give us a sense whether or not the theory derived above matches up with reality.

The TAQ quote and transaction data is obtained for Alcoa (AA), Altria Group (MO), Citigroup (C), General Electric (GE), and International Business Machines (IBM) over the period January 2, 2004 through December 31, 2004 (252 trading days). Following Hansen and Lunde (2006), we only consider data for the main NYSE market during the time interval 9.45 – 16.00. For quotes, we apply a filtering algorithm that selects an observation if it satisfies the following conditions (i) the bid price and / or the offer price are improved relative to the prevailing best quote and (ii) the spread between the offer and the bid is less than $\$1$. Because the securities we consider are very liquid, spreads in excess of one dollar are rare, and if they do occur often indicate a recording error or unreliable quote. The first condition filters out quotes that are wide and those which are the same as the best prevailing quote but with different volume (we refer to the latter as "liquidity quotes").

Table 1 present some summary statistics of the data and gives details of the quote filtering. A couple of points are worth highlighting. First, about 90% of the raw quote data constitute liquidity quotes that reinforce the best available quote by altering its volume. Only a small proportion of the quotes is uncompetitive. However, since we cannot keep track of quote deletions, the wide quotes could in fact be competitive quotes when the best available quote is withdrawn. Given their relatively infrequent occurrence, it is unlikely that this filtering will have a substantive impact on the results. Second, the magnitude of serial correlation in both the transaction data and the mid-quote data is relatively modest. This is due to the fact that we only consider NYSE data, thereby avoiding potential contamination of "noisy" quotes/transactions from satellite markets. Still, for transactions the first order serial correlation ranges from -15% for IBM, AA, and MO to -35% for GE. Second and higher order correlations are much smaller. For quotes we observe a similar pattern with the only exception that second order correlations are more sizeable and consistently positive. Because the quote data are sampled only when revisions occur, this is in essence equivalent to tick sampling and thereby ensures that the long sequences of "zero-returns" commonly found in transaction data do not arise. As discussed in Griffin and Oomen (2005), tick sampling leads to substantial high order serial correlation with alternating sign. This is consistent with the results presented here.

Turning to the covariance signatures, Figures 7 and 8 plot the daily RC and RCLL(1,1) measures, averaged across days, for sampling frequencies ranging from 1 second ($M = 22500$) to 5 minutes ($M = 75$). The dashed lines indicate the theoretical signatures based on Eqs. 5 and 7 above. Here, estimates for λ are simply obtained as the average number of observations per day for each asset (see Table 1), and estimates for $\rho_{ij}\sigma_i\sigma_j$ are computed from the 5 minute data. For ease of presentation, all graphs are rescaled. Given the relatively stylized setting in which the theoretical results have been derived, the correspondence between empirical and theoretical signature plots is striking. The rate at which the RC decays with an increase in sampling frequency matches up almost perfectly with that predicted by the simple model with independent Poisson sampling. Also for RC with an ad hoc first order lead and lag adjustment, the results look good. Of course, deviations can be expected but it is difficult to spot any systematic discrepancies in these graphs.

Table 1: Summary statistics of NYSE quote and transaction data 2004

	AA	MO	C	GE	IBM
Panel A: quote data					
Raw quotes	2,448,107	2,521,171	3,230,580	3,587,421	3,276,757
liquidity	2,148,380	2,173,275	2,855,562	3,377,937	2,617,084
wide	67,398	74,871	56,083	29,246	159,298
spread$>$ \$1	22	10	0	0	0
# Quotes	232,318	273,024	318,935	180,238	500,375
Quote intensity λ	922	1,083	1,265	715	1,986
Correlation ρ_1	–10.1	–10.5	–14.7	–16.3	–11.1
ρ_2	7.46	6.11	5.37	10.0	3.95
ρ_3	1.30	0.62	–2.36	–4.21	0.22
Noise ratio γ	0.109	0.110	0.305	0.339	0.190
Average spread (cents)	1.54	1.69	1.40	1.21	2.08
Panel B: transaction data					
# Trades	834,857	927,783	1,329,289	1,333,754	1,230,409
Trade intensity λ	3,313	3,682	5,275	5,293	4,883
Correlation ρ_1	–17.4	–16.8	–23.3	–35.0	–14.8
ρ_2	–1.32	–2.41	–2.18	–1.76	–3.55
ρ_3	1.86	2.06	0.97	1.26	0.46
Noise ratio γ	0.358	0.375	0.561	0.905	0.356

Note. NYSE quotes and transactions between 9.45 and 16.00 from January 2, 2004 through to December 31, 2004. Noise ratio measures level of market microstructure noise (see Oomen, 2006, for more details). Intensity estimates are equal to number of quotes / transactions divided by total number of days in sample period.

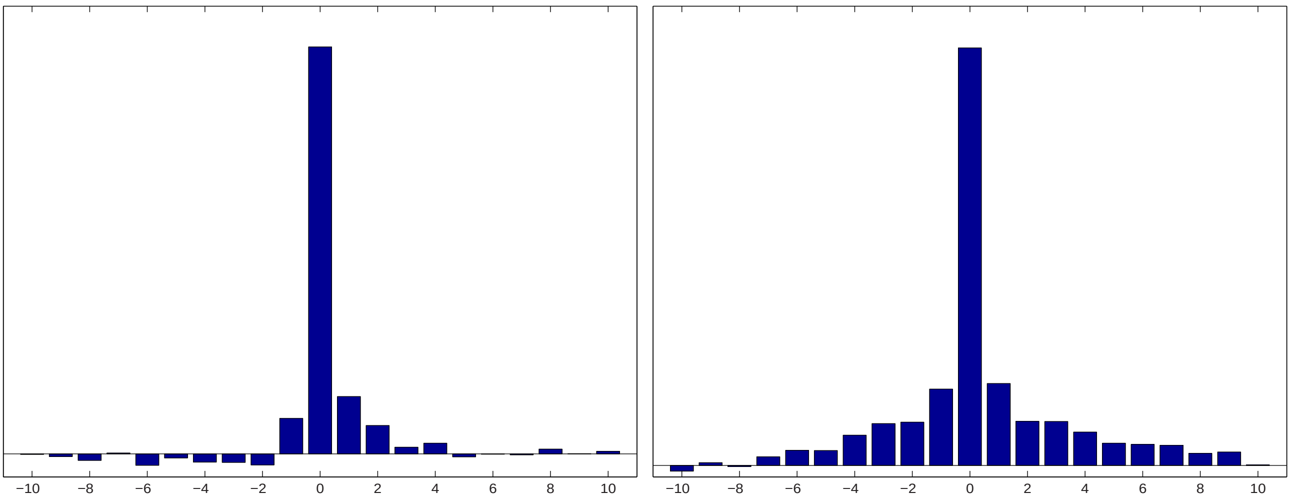
Figure 9 presents the signature plots for the HY estimator. Here the daily HY estimator, averaged across days, is computed as a function of k (to facilitate presentation, all graphs are rescaled to ensure that the average lies at 1). Of course, from the theory we expect no significant departures in the HY estimates as the aggregation frequency is varied because it's expectation has been shown to be unaffected by non-synchronicity and i.i.d. noise. The empirical results, however, indicate that this is not the case at the very highest frequencies. In fact, a sizable systematic downward bias of about 20% is detected in the HY estimator computed with quote-to-quote mid prices. Interestingly, sampling every 2–3 quotes seems to eliminate this bias.

Figures 10, 11, and 12 present the corresponding results for the transaction data. A similar pattern to the quote data emerges with two important differences. First, the decay of the RC and RCLL signature plots appears

Figure 6: IBM return dependence with leads & lags of GE returns (2004)

Panel A: quote-update returns

Panel B: transaction returns



Note. This figure plots $\phi(h)$ as a function of h for IBM and GE quote-update returns (Panel A) and transaction returns (Panel B).

somewhat faster which we suspect has to do with the relatively frequent occurrence of zero returns in transaction data (see Griffin and Oomen, 2005), a feature that is clearly not captured in our framework. Second, the bias of the HY estimator is substantially larger (about 50%!) and is only eliminated after aggregating returns to a frequency of about 10 transactions.

To gain some further insights into the bias of HY, consider the statistic $\phi(h)$ defined as $\sum_i R_i^{(1)} R_{\max A_i+h}^{(2)}$ for $h > 0$ and $\sum_i R_i^{(1)} R_{\min A_i+h}^{(2)}$ for $h < 0$ and $\sum_i \sum_{j \in A_i} R_i^{(1)} R_j^{(2)}$ for $h = 0$ (where A_i is as defined in Eq. 11). Intuitively, $\phi(h)$ measures the sample covariance of quote/transaction returns between asset 1 and 2. Note that in the modeling framework adopted here, $E[\phi(0)] = E[HY] = \rho\sigma_1\sigma_2$, and $E[\phi(h)] = 0$ for $h \neq 0$ because leads and lags of returns that share no overlap in time carry no information about the underlying correlation structure. This is precisely the reason why HY is unbiased in theory. However, cross dependence between non-overlapping returns can arise in practice when price adjustment is not instantaneous and it takes some trading before prices fully reflect the currently available information. In such a scenario, one would of course expect HY to be biased because it misses out on cross dependence that extends beyond the overlapping transaction or quote returns. To illustrate that this may be a reasonable conjecture, Figure 6 plots $\phi(h)$ averaged day-by-day over 2004 for IBM and GE quote and transaction returns in Panels A and B respectively. The lead-lag dependence is evident and the pattern appears much stronger for transaction returns than for quote returns. Based on Figure 6 we can derive an "implied bias" for the HY estimator as $1-\phi(0)/\sum_h \phi(h)$. Doing so for IBM and GE, using the first 10 leads and lags, we calculate a bias of 18.0% for quotes and 55.4% for transactions which is very much in line with the bias reported in Figures 9 and 12. So all in all, we find that at least part of the observed HY bias can be accounted for by cross dependence of non-overlapping returns that arise from "sluggish" adjustment of quotes and transaction prices.

Given the above, it seems natural to modify the HY estimator as follows (ignoring end-effects):

$$HYLL = \sum_{i=1}^{M_1} \sum_{j=\min A_i - L}^{\max A_i + U} R_i^{(1)} R_j^{(2)}. \tag{15}$$

This estimator can of course be further extended to include kernel weights. Setting U and L sufficiently high would eliminate the bias at the cost of a higher variance, paralleling the case for RCLL discussed above. However, an obvious drawback of HYLL (one that RCLL importantly does not suffer from) is that because it is implemented in "event time" it is not invariant to the ordering of assets, meaning that covariance estimates between asset 1 and 2 will not be the same as those between asset 2 and 1, even after appropriately switching U and L. While a detailed study of the lead-lag return dependence and extensions of the HY estimator are clearly of interest, this would necessitate more complicated models that bring us outside the realm of arbitrage free pricing and, as such, pose a substantial challenge that we do not attempt to address in this paper.

4 Conclusion

This paper studies the statistical properties of three popular covariance estimators, namely realized covariance (RC), realized covariance plus lead-lag adjustments (RCLL), and the Hayashi-Yoshida (HY) estimator, in a setting where prices are observed with noise and non-synchronously in time. We derive closed form bias and variance expressions for all estimators and use these to provide a detailed discussion of their relative efficiency. The main finding of this paper is that the ordering of the competing covariance estimators in terms of their efficiency is primarily determined by the level of microstructure noise. The empirical results indicate that the rate of decay in the covariance signature plots is roughly consistent with that implied by our theory. Interestingly, the

HY estimator is severely downward biased at the highest sampling frequencies and we present some evidence that this is probably caused by “sluggish” adjustment of prices. Various kernel-based and subsampling extensions of the RC and HY estimators are suggested but a formal analysis of these is left for future research.

A Proofs

Let $t_i^{*(h)}$ denote the timing of the most recent observation (i.e. transaction / quote-update) of asset h prior to $t = i\Delta$. Define $\tau_i^{(h)} = i\Delta - t_i^{*(h)}$ and $a_i = \max\{\tau_i^{(1)}, \tau_i^{(2)}\}$ and $b_i = \min\{\tau_i^{(1)}, \tau_i^{(2)}\}$. Let $E_i^{(h)}$ denote the event that asset h trades at least once over the time interval $((i-1)\Delta, i\Delta]$. Let $\nu\left(I_1 \cap I_2\right)$ denotes the length of the intersection between interval I_1 and I_2. In sections A.1 and A.2 below we also use the following short-hands $R_i = r_i^{(1)}$, $Z_i = r_i^{(2)}$, $u_i = u_{N_1(i\Delta)}^{(1)}$, $v_i = u_{N_2(i\Delta)}^{(2)}$ and $\mu_h = 1 - e^{-\lambda_h \Delta}$ for simplicity of notation.

A.1 Proof of Theorem 2.1

To calculate $E\left(R_i Z_i\right)$ we condition on a trade of asset R and Z in interval $\left((i-1)\Delta, i\Delta\right]$.

$$\begin{aligned}
E\left(R_i Z_i\right) &= \Pr\{E_i^{(1)} \cap E_i^{(2)}\} E(R_i Z_i | E_i^{(1)} \cap E_i^{(2)}) \\
&= \mu_1 \mu_2 \rho \sigma_1 \sigma_2 E(\nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{i-1}^{(2)}, t_i^{(2)})) | E_i^{(1)} \cap E_i^{(2)}) \\
&= \mu_1 \mu_2 \rho \sigma_1 \sigma_2 E\left(b_{i-1} + \Delta - a_i | a_i < \Delta\right) \\
&= \rho \sigma_1 \sigma_2 \left(\Delta - \frac{\lambda_1^2 \mu_2 + \lambda_2^2 \mu_1}{\lambda_1 \lambda_2 \left(\lambda_1 + \lambda_2\right)}\right).
\end{aligned}$$

To derive $E(b_i)$ and $E(a_i|a_i < \Delta)$, we use that for exponential variable z_i with mean λ_i^{-1} the following holds:

$$\begin{aligned}
F_{\min}\left(u\right) &= \Pr\left(\min\left\{z_1, z_2\right\} < u\right) = 1 - \Pr\left(z_1 > u \cap z_2 > u\right) = 1 - e^{-(\lambda_1 + \lambda_2) u}, \\
F_{\max}\left(u\right) &= \Pr\left(\max\left\{z_1, z_2\right\} < u\right) = \Pr\left(z_1 < u \cap z_2 < u\right) = \left(1 - e^{-\lambda_1 u}\right)\left(1 - e^{-\lambda_2 u}\right).
\end{aligned}$$

Thus,

$$E\left(b_i\right) = \int_0^\infty u\left(\lambda_1 + \lambda_2\right) e^{-(\lambda_1 + \lambda_2) u} du = \frac{1}{\lambda_1 + \lambda_2},$$

and $\kappa_1 \equiv E\left(a_i | a_i < \Delta\right)$ is equal to:

$$\begin{aligned}
\kappa_1 &= \frac{1}{\mu_1 \mu_2} \int_0^\Delta u \left(\lambda_1 e^{-u\lambda_1} + \lambda_2 e^{-u\lambda_2} - \left(\lambda_1 + \lambda_2\right) e^{-(\lambda_1 + \lambda_2) u}\right) du \\
&= \Delta - \frac{\Delta}{\mu_1 \mu_2} + \frac{1}{\lambda_1 + \lambda_2} \left(\frac{\lambda_1}{\mu_1 \lambda_2} + \frac{\lambda_2}{\mu_2 \lambda_1} + 1\right). \qquad (16)
\end{aligned}$$

The expression in Eq. (5) now follows from $E\left(RC\right) = ME\left(R_i Z_i\right)$. To prove unbiasedness in presence of noise, simply note that $E\left(R_i + u_i + u_{i-1}\right)\left(Z_i + v_i + v_{i-1}\right) = E\left(R_i Z_i\right)$ due to the i.i.d. noise assumption.

To work out the variance of RC proceed as follows:

$$\left(RC_M\right)^2 = \sum_{i=1}^M \sum_{j=1}^M R_i Z_i R_j Z_j.$$

First consider the case where $i = j$. If we condition on $E_i^{(1)} \cap E_i^{(2)}$ and the transaction arrival times then returns are jointly normal and so we have:

$$\begin{aligned}
E_t\left(R_i^2 Z_i^2\right) &= \Pr\{E_i^{(1)} \cap E_i^{(2)}\}(E_t(R_i^2 | E_i^{(1)}) E_t(Z_i^2 | E_i^{(2)}) + 2\rho^2 E_t(R_i Z_i | E_i^{(1)} \cap E_i^{(2)})^2) \\
&= \mu_1 \mu_2 \sigma_1^2 \sigma_2^2 ((t_i^{*(1)} - t_{i-1}^{*(1)})(t_i^{*(2)} - t_{i-1}^{*(2)}) + 2\rho^2 \left(b_{i-1} + \Delta - a_i\right)^2).
\end{aligned}$$

Now taking expectations over t we get:

$$\begin{aligned} E\left(R_i^2 Z_i^2\right) &= \sigma_1^2\sigma_2^2\Delta^2 + 2\mu_1\mu_2\rho^2\sigma_1^2\sigma_2^2 E\left(b_{i-1}^2 + 2b_{i-1}\Delta - 2b_{i-1}a_i + \Delta^2 - 2\Delta a_i + a_i^2 | a_i < \Delta\right) \\ &= \sigma_1^2\sigma_2^2\left(1+2\rho^2\right)\Delta^2 + 4\frac{\rho^2\sigma_1^2\sigma_2^2}{\lambda_1+\lambda_2}\left(\frac{\mu_1\mu_2}{\lambda_1+\lambda_2} + \frac{\lambda_1\mu_2}{\lambda_2^2} + \frac{\lambda_2\mu_1}{\lambda_1^2} - \left(\frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1}\right)\Delta\right), \end{aligned}$$

using that $E(a_{i-1}b_i|a_{i-1} < \Delta) = E(a_{i-1}|a_{i-1} < \Delta)E(b_i)$, $E\left(b_i^2\right) = 2\left(\lambda_1+\lambda_2\right)^{-2}$ and $\kappa_2 \equiv E\left(a_i^2|a_i < \Delta\right)$ with

$$\kappa_2 = \Delta^2 - \frac{\Delta^2}{\mu_1\mu_2} + 2\frac{\mu_1 - \lambda_1\Delta e^{-\lambda_1\Delta}}{\mu_1\mu_2\lambda_1^2} + 2\frac{\mu_2 - \lambda_2\Delta e^{-\lambda_2\Delta}}{\mu_1\mu_2\lambda_2^2} + 2\frac{e^{-\Delta(\lambda_1+\lambda_2)}\Delta}{\mu_1\mu_2\left(\lambda_1+\lambda_2\right)} + 2\frac{e^{-\Delta(\lambda_1+\lambda_2)} - 1}{\mu_1\mu_2\left(\lambda_1+\lambda_2\right)^2}. \tag{17}$$

To work out the off-diagonal elements (i.e. $i \neq j$), condition on the trade arrival times:

$$E_t\left[Z_iR_iZ_jR_j\right] = E_t\left[Z_iR_i\right]E_t\left[Z_jR_j\right] + E_t\left[Z_iZ_j\right]E_t\left[R_iR_j\right] + E_t\left[R_iZ_j\right]E_t\left[Z_iR_j\right].$$

The second and third term are zero. The first term can be written as:

$$E\left(E_t\left[Z_iR_i\right]E_t\left[Z_jR_j\right]|F_1\right) = \left(E\left(Z_iR_i\right)\right)^2 + \delta. \tag{18}$$

In section A.1.1 below we show that δ is negligible and thus we ignore it for simplicity of exposition. Collecting all terms, the variance of RC is now equal to:

$$V\left(RC_M\right) = \sigma_1^2\sigma_2^2\frac{1+2\rho^2-\rho^2\beta_M^2}{M} - 4\frac{\rho^2\sigma_1^2\sigma_2^2}{\lambda_1+\lambda_2}\left(\frac{\lambda_1}{\lambda_2}+\frac{\lambda_2}{\lambda_1}\right) + 4M\frac{\rho^2\sigma_1^2\sigma_2^2}{\lambda_1+\lambda_2}\left(\frac{\mu_1\mu_2}{\lambda_1+\lambda_2} + \frac{\lambda_1\mu_2}{\lambda_2^2} + \frac{\lambda_2\mu_1}{\lambda_1^2}\right),$$

which corresponds to Eq. (6) with $\xi_1 = \xi_2 = 0$.

With i.i.d. noise let the contaminated returns be denoted as $R_i + u_i - u_{i-1}$ and $Z_i + v_i - v_{i-1}$. Then, the noise contribution to the variance of RC can be expressed as follows:

$$\begin{aligned} V(RC_M^{(iid)}) - V(RC_M) &= E\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\left(R_i+u_i-u_{i-1}\right)\left(Z_i+v_i-v_{i-1}\right)\left(R_j+u_j-u_{j-1}\right)\left(Z_j+v_j-v_{j-1}\right) - R_iZ_iR_jZ_j\right) \\ &= \mu_1\mu_2\sum_{i=1}^{M}E(2v_i^2R_i^2 + 2u_i^2Z_i^2 + 4u_i^2v_i^2|E_i^{(1)}\cap E_i^{(2)}) + 2\mu_1^2\mu_2^2\sum_{i=1}^{M-1}E(u_i^2v_i^2|E_{i,i+1}^{(1)}\cap E_{i,i+1}^{(2)}) \\ &= 2\mu_2\sigma_1^2\xi_2^2 + 2\mu_1\sigma_2^2\xi_1^2 + 4M\mu_1\mu_2\xi_1^2\xi_2^2 + 2\left(M-1\right)\mu_1^2\mu_2^2\xi_1^2\xi_2^2, \end{aligned}$$

using that $E(v_i^2R_i^2|E_i^{(1)}\cap E_i^{(2)}) = \xi_2^2\sigma_1^2\Delta/\mu_1$. Collecting terms gives the required expression in Eq. (6). ■

A.1.1 A note on the approximation in Eq. (18)

Let $F_1 \equiv E_i^{(1)}\cap E_i^{(2)}\cap E_j^{(1)}\cap E_j^{(2)}$. Assuming $j > i$, we have:

$$\begin{aligned} E[E_t[Z_iR_i]E_t[Z_jR_j]] &= \rho^2\sigma_1^2\sigma_2^2\mu_1^2\mu_2^2E[\left(b_{i-1}+\Delta-a_i\right)\left(b_{j-1}+\Delta-a_j\right)|F_1] \\ &= \rho^2\sigma_1^2\sigma_2^2\mu_1^2\mu_2^2E[\Delta^2 - a_j\Delta + b_{i-1}\Delta - a_i\Delta - b_{i-1}a_j + a_ia_j|F_1] \\ &\quad +\rho^2\sigma_1^2\sigma_2^2\mu_1^2\mu_2^2\left(E[b_{i-1}+\Delta|F_1]E[b_{j-1}|F_1] - E[a_ib_{j-1}|F_1]\right). \end{aligned}$$

Using previous results, the first term is easily worked out as:

$$E[\Delta^2 - a_j\Delta + b_{i-1}\Delta - a_i\Delta - b_{i-1}a_j + a_ia_j|F_1] = \left(\Delta-\kappa_1\right)^2 + \frac{\Delta-\kappa_1}{\lambda_1+\lambda_2}.$$

Similarly,

$$E[\Delta + b_{i-1}|F_1] = \Delta + \frac{1}{\lambda_1 + \lambda_2}.$$

The non–trivial terms are those involving b_{j-1} because conditional on F_1, $b_{j-1} < (j-i)\,\Delta$ and if $b_{j-1} > (j-i-1)\,\Delta$ then b_{j-1} is not independent of a_i. First, consider the case where $b_{j-1} < (j-i-1)\,\Delta$. Then:

$$E\left[b_{j-1}|b_{j-1} < (j-i-1)\,\Delta, F_1\right] = \frac{1 - e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\left((\lambda_1+\lambda_2)(j-i-1)\,\Delta + 1\right)}{(\lambda_1+\lambda_2)\left(1 - e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\right)}.$$

Also, we have:

$$\begin{aligned}
& E\left[b_{j-1}|\left(j-i-1\right)\Delta < b_{j-1} < (j-i)\,\Delta, F_1\right] \\
= \;& (j-i-1)\,\Delta + E\left[b_i|a_i < \Delta\right] \\
= \;& (j-i-1)\,\Delta + \frac{\lambda_1\lambda_2}{\mu_1\mu_2}\left(\int_0^{\Delta}\int_0^{z_2} z_1 e^{-z_1\lambda_1 - z_2\lambda_2} dz_1 dz_2 + \int_0^{\Delta}\int_0^{z_1} z_2 e^{-z_1\lambda_1 - z_2\lambda_2} dz_2 dz_1\right) \\
= \;& (j-i-1)\,\Delta + \frac{1}{\lambda_1+\lambda_2} + \frac{e^{-\Delta(\lambda_1+\lambda_2)}\left(\lambda_1\lambda_2^2\Delta + \lambda_2^2 + \lambda_1^2\lambda_2\Delta + \lambda_1^2\right) - \lambda_1^2 e^{-\lambda_1\Delta} - \lambda_2^2 e^{-\lambda_2\Delta}}{\mu_1\mu_2\lambda_1\lambda_2\,(\lambda_1+\lambda_2)}.
\end{aligned}$$

Combining the above, we get:

$$\begin{aligned}
E\left[b_{j-1}|F_1\right] &= E\left[b_{j-1}|b_{j-1} < (j-i-1)\,\Delta, F_1\right]\left(1 - e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\right) \\
&\quad + E\left[b_{j-1}|\left(j-i-1\right)\Delta < b_{j-1} < (j-i)\,\Delta, F_1\right] e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta} \\
&= \frac{1}{\lambda_1+\lambda_2} + e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\kappa_3,
\end{aligned}$$

where

$$\kappa_3 = \frac{1}{\mu_1\mu_2}\left(e^{-(\lambda_1+\lambda_2)\Delta}\Delta - \frac{\lambda_1}{\lambda_2}\frac{\mu_2 e^{-\lambda_1\Delta}}{\lambda_1+\lambda_2} - \frac{\lambda_2}{\lambda_1}\frac{\mu_1 e^{-\lambda_2\Delta}}{\lambda_1+\lambda_2}\right). \tag{19}$$

To work out the second term, we distinguish among the same cases, namely $b_{j-1} < (j-i-1)\,\Delta$ and $(j-i-1)\,\Delta < b_{j-1} < (j-i)\,\Delta$.

$$\begin{aligned}
E\left[a_i b_{j-1}|b_{j-1} < (j-i-1)\,\Delta, F_1\right] &= E\left[a_i|F_1\right] E\left[b_{j-1}|b_{j-1} < (j-i-1)\,\Delta, F_1\right] \\
&= \kappa_1\frac{1 - e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\left((\lambda_1+\lambda_2)(j-i-1)\,\Delta + 1\right)}{(\lambda_1+\lambda_2)\left(1 - e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\right)},
\end{aligned}$$

and

$$E\left[a_i b_{j-1}|\left(j-i-1\right)\Delta < b_{j-1} < (j-i)\,\Delta, F_1\right] = \kappa_1\,(j-i-1)\,\Delta + E\left[a_i b_i|F_1\right],$$

where

$$\begin{aligned}
\kappa_4 &= E\left[a_i b_i|F_1\right] = \frac{1}{\mu_1\mu_2}\int_0^{\Delta}\int_0^{\Delta} z_1 z_2 \lambda_1 e^{-z_1\lambda_1}\lambda_2 e^{-z_2\lambda_2} dz_1 dz_2 \\
&= \frac{\left(\mu_2 - \lambda_2 e^{-\lambda_2\Delta}\Delta\right)\left(\mu_1 - \lambda_1 e^{-\lambda_1\Delta}\Delta\right)}{\mu_1\mu_2\lambda_1\lambda_2}.
\end{aligned} \tag{20}$$

Combining the above terms we have:

$$\begin{aligned}
E\left[a_i b_{j-1}|F_1\right] &= E\left[a_i b_{j-1}|b_{j-1} < (j-i-1)\,\Delta, F_1\right]\left(1 - e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\right) \\
&\quad + E\left[a_i b_{j-1}|\left(j-i-1\right)\Delta < b_{j-1} < (j-i)\,\Delta, F_1\right] e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta} \\
&= \frac{\kappa_1}{\lambda_1+\lambda_2} + e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\left(\kappa_4 - \frac{\kappa_1}{\lambda_1+\lambda_2}\right).
\end{aligned}$$

Thus:

$$\begin{aligned} E[E_t[Z_iR_i]E_t[Z_jR_j]] &= \rho^2\sigma_1^2\sigma_2^2\mu_1^2\mu_2^2\left(\frac{\left(1+\left(\Delta-\kappa_1\right)\left(\lambda_1+\lambda_2\right)\right)^2}{\left(\lambda_1+\lambda_2\right)^2}+\left(\kappa_3\Delta-\kappa_4+\frac{\kappa_3+\kappa_1}{\lambda_1+\lambda_2}\right)e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}\right) \\ &= \rho^2\sigma_1^2\sigma_2^2\left(\Delta-\frac{\lambda_1^2\mu_2+\lambda_2^2\mu_1}{\lambda_1\lambda_2\left(\lambda_1+\lambda_2\right)}\right)^2-\left(\frac{\rho\sigma_1\sigma_2\mu_1\mu_2}{\lambda_1+\lambda_2}\right)^2 e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta} \\ &= \left(E\left(Z_iR_i\right)\right)^2-\left(\frac{\rho\sigma_1\sigma_2\mu_1\mu_2}{\lambda_1+\lambda_2}\right)^2 e^{-(\lambda_1+\lambda_2)(j-i-1)\Delta}. \end{aligned} \tag{21}$$

In the summation over $j \neq i$ up to M, the first term is of order M^2 whereas the second term is of order M and negative. Hence, $E[E_t[Z_jR_j]E_t[Z_iR_i]]$ is very accurately bounded by $\left(E\left(Z_iR_i\right)\right)^2$.

A.2 Proof of Theorem 2.2

To calculate $E\left(R_{i-h}Z_i\right)$ we condition on $E_{i-h}^{(1)} \cap E_i^{(2)}$ plus no trade for asset Z over the interval $[(i-h-1)\Delta, (i-1)\Delta]$. The expectation can then be expressed as:

$$\begin{aligned} E\left(R_{i-h}Z_i\right) &= \rho\sigma_1\sigma_2\mu_1\mu_2 e^{-\lambda_2(h-1)\Delta}E(\tau_{i-h}^{(2)}-\tau_{i-h}^{(1)}|\tau_{i-h}^{(1)}<\tau_{i-h}^{(2)}<\Delta+\tau_{i-h-1}^{(1)}) \\ &\quad +\rho\sigma_1\sigma_2\mu_1\mu_2 e^{-\lambda_2(h-1)\Delta}E(\tau_{i-h-1}^{(1)}+\Delta-\tau_{i-h}^{(1)}|\tau_{i-h}^{(2)}>\Delta+\tau_{i-h-1}^{(1)}) \\ &= \rho\sigma_1\sigma_2\frac{\lambda_1\left(1-e^{-\lambda_2\Delta}\right)^2}{\lambda_2(\lambda_1+\lambda_2)}e^{-\lambda_2(h-1)\Delta}, \end{aligned}$$

using that τ is exponential and

$$\begin{aligned} E\left(w_2-z_1|z_1<w_2<\Delta+w_1\right) &= \int_0^\infty\int_0^\Delta\int_{z_1}^{\Delta+w_1}\left(w_2-z_1\right)\frac{\lambda_1 e^{-\lambda_1 z_1}\lambda_1 e^{-\lambda_1 w_1}\lambda_2 e^{-\lambda_2 w_2}}{1-e^{-\lambda_1\Delta}}dw_2dz_1dw_1 \\ E\left(w_1+\Delta-z_1|w_2>\Delta+w_1\right) &= \int_0^\infty\int_0^\Delta\int_{\Delta+w_1}^\infty\left(w_2-z_1\right)\frac{\lambda_1 e^{-\lambda_1 z_1}\lambda_1 e^{-\lambda_1 w_1}\lambda_2 e^{-\lambda_2 w_2}}{1-e^{-\lambda_1\Delta}}dw_2dz_1dw_1, \end{aligned}$$

where $w_1, z_1 \sim Exp(1/\lambda_1)$, $w_2 \sim Exp(1/\lambda_2)$ for $0<z_1<\Delta$ and $0<w_1,w_2<\infty$. By symmetry we have

$$E\left(R_{i+h}Z_i\right)=\rho\sigma_1\sigma_2\frac{\lambda_2\left(1-e^{-\lambda_1\Delta}\right)^2}{\lambda_1(\lambda_1+\lambda_2)}e^{-\lambda_1(h-1)\Delta}.$$

Using the above, the expectation of RCLL in Eq. (8) directly follows.

To work out the variance of RCLL proceed as follows:

$$(RCLL)^2=\sum_{i=1}^M\sum_{j=1}^M\overline{R}_iZ_i\overline{R}_jZ_j,$$

where $\overline{R}_i=\sum_{l=-L}^U R_{i+l}$. Below we redefine $E_i^{(1)}$ to denote the conditioning event where there is a trade of asset R in the interval $[(i-L-1)\Delta,(i+U)\Delta]$. First consider the case where $i=j$. If we condition on $E_i^{(1)}\cap E_i^{(2)}$ and the transaction arrival times, then returns are jointly normal and we have:

$$\begin{aligned} E_t[\overline{R}_i^2Z_i^2] &= \Pr\{E_i^{(1)}\cap E_i^{(2)}\}(E_t(\overline{R}_i^2|E_i^{(1)})E_t(Z_i^2|E_i^{(2)})+2\rho^2(E_t(\overline{R}_iZ_i|E_i^{(1)}\cap E_i^{(2)}))^2) \\ &= \sigma_1^2\sigma_2^2\mu_1^*\mu_2(t_{i+U}^{*(1)}-t_{i-L-1}^{*(1)})(t_i^{*(2)}-t_{i-1}^{*(2)}) \\ &\quad +2\rho^2\sigma_1^2\sigma_2^2\mu_1^*\mu_2(\max[0,\min\{t_{i+U}^{*(1)},t_i^{*(2)}\}-\max\{t_{i-L-1}^{*(1)},t_{i-1}^{*(2)}\}])^2. \end{aligned}$$

Now taking expectations over t, the first term on the right hand side is:

$$E[\sigma_1^2\sigma_2^2\mu_1^*\mu_2(t_{i+U}^{*(1)} - t_{i-L-1}^{*(1)})(t_i^{*(2)} - t_{i-1}^{*(2)})] = \sigma_1^2\sigma_2^2\left(U+L+1\right)\Delta^2.$$

To simplify notation, define $z_1 = \tau_{i+U}^{(1)}, w_1 = \tau_{i-L-1}^{(1)}, z_2 = \tau_i^{(2)}$, and $w_2 = \tau_{i-1}^{(2)}$. To work out the expectation of the second term, distinguish among following four conditioning states:

$$\begin{aligned}
A_1 &: \quad 0 < z_1 < U\Delta \\
A_2 &: \quad U\Delta < z_1 < \left(U+1\right)\Delta \\
A_3 &: \quad \left(U+1\right)\Delta < z_1 < \min\left\{\left(U+1\right)\Delta + w_2, \left(U+L+1\right)\Delta\right\} \\
A_4 &: \quad \min\left\{\left(U+1\right)\Delta + w_2, \left(U+L+1\right)\Delta\right\} < z_1 < \left(U+L+1\right)\Delta
\end{aligned}$$

To calculate the state probabilities we use that $\{z_1, w_1, z_2, w_2\}$ are jointly exponential and in particular that $p(z_1, w_2) = \lambda_1\lambda_2\exp\{-\lambda_1 z_1 - \lambda_2 w_2\}/\mu_1^*$ for $0 < z_1 < \left(U+L+1\right)\Delta$ and $0 < w_2 < \infty$.

$$\begin{aligned}
\Pr\left\{A_1\right\} &= (1 - e^{-U\lambda_1\Delta})/\mu_1^* \\
\Pr\left\{A_2\right\} &= \mu_1 e^{-U\lambda_1\Delta}/\mu_1^* \\
\Pr\left\{A_3\right\} &= \frac{\lambda_1}{\lambda_1+\lambda_2}(1 - e^{-L(\lambda_1+\lambda_2)\Delta})e^{-(U+1)\lambda_1\Delta}/\mu_1^* \\
\Pr\left\{A_4\right\} &= \frac{\lambda_1 e^{-L\Delta(\lambda_1+\lambda_2)} + \lambda_2}{\left(\lambda_1+\lambda_2\right)\mu_1^*}e^{-(U+1)\Delta\lambda_1} - \frac{1-\mu_1^*}{\mu_1^*}
\end{aligned}$$

Note that when $U = L = 0$ then $\Pr\left\{A_2\right\} = 1$ and when $L = 0$ then $\Pr\left\{A_4\right\} = 0$. Also in state A_4, $\max\{0, \min\{t_{i+U}^{*(1)}, t_i^{*(2)}\} - \max\{t_{i-L-1}^{*(1)}, t_{i-1}^{*(2)}\}\} = 0$ and so to work out the expectation we only need to consider the first three states.

In case A_1 we have:

$$\min\{t_{i+U}^{*(1)}, t_i^{*(2)}\} - \max\{t_{i-L-1}^{*(1)}, t_{i-1}^{*(2)}\} = \Delta - z_2 + B$$

with z_2 independent of $B = \min\left\{w_2, L\Delta + w_1\right\}$.

$$\begin{aligned}
E[(\Delta - z_2 + B)^2] &= \Delta^2 - 2\Delta E\left(z_2\right) + 2\Delta E\left(B\right) + E\left(z_2^2\right) - 2E\left(z_2\right)E\left(B\right) + E\left(B^2\right) \\
&= \frac{2}{\lambda_2^2} + \frac{\Delta^2}{\mu_2} - 2\lambda_1\frac{\Delta\left(L\mu_2+1\right)\left(\lambda_1+\lambda_2\right) + \mu_2}{\lambda_2\left(\lambda_1+\lambda_2\right)^2\mu_2}e^{-\lambda_2 L\Delta} \qquad (22)
\end{aligned}$$

using that

$$\begin{aligned}
E\left(z_2 | z_2 < \Delta\right) &= \frac{1 - \left(1+\lambda_2\Delta\right)e^{-\lambda_2\Delta}}{\lambda_2\left(1 - e^{-\lambda_2\Delta}\right)} \\
E\left(z_2^2 | z_2 < \Delta\right) &= \frac{2}{\lambda_2^2} - \Delta\frac{\left(2+\lambda_2\Delta\right)e^{-\lambda_2\Delta}}{\lambda_2\left(1 - e^{-\lambda_2\Delta}\right)} \\
E\left(B\right) &= \frac{1}{\lambda_2} - \frac{\lambda_1 e^{-\lambda_2 L\Delta}}{\lambda_2\left(\lambda_2+\lambda_1\right)} \\
E\left(B^2\right) &= \frac{2}{\lambda_2^2} - 2\frac{\lambda_1\left(\lambda_2 L\Delta\left(\lambda_2+\lambda_1\right) + \lambda_1 + 2\lambda_2\right)e^{-\lambda_2 L\Delta}}{\lambda_2^2\left(\lambda_2+\lambda_1\right)^2}
\end{aligned}$$

and the distribution of $B = \min\left\{w_2, L\Delta + w_1\right\}$ is given as:

$$f_B\left(b\right) = \begin{cases} \lambda_2 e^{-\lambda_2 b} & b < L\Delta \\ \left(\lambda_2+\lambda_1\right)e^{-\lambda_2 L\Delta}e^{-(\lambda_1+\lambda_2)(b-L\Delta)} & b > L\Delta \end{cases}$$

In case A_2 we have:

$$\min\{t_{i+U}^{*(1)}, t_i^{*(2)}\} - \max\{t_{i-L-1}^{*(1)}, t_{i-1}^{*(2)}\} = \Delta - A + B$$

where $\widetilde{z}_1 = z_1 - \Delta U$ and $A = \max\{\widetilde{z}_1, z_2\}$ is independent of $B = \min\{w_2, L\Delta + w_1\}$.

$$\begin{aligned} E[(\Delta - A + B)^2] &= \Delta^2 - 2\Delta E(A) + 2\Delta E(B) + E(A^2) - 2E(A)E(B) + E(B^2) \\ &= \frac{\Delta^2}{\mu_1\mu_2} + \frac{2}{\lambda_1+\lambda_2}\frac{1}{\mu_1\mu_2}\left(\frac{\mu_1\mu_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_2^2}\mu_2 + \frac{\lambda_2}{\lambda_1^2}\mu_1 - \Delta\left(\frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1}\right)\right) \\ &\quad +2\frac{\lambda_1(\lambda_1+\lambda_2)(\mu_1\mu_2+\lambda_2\Delta) - \lambda_1^2\mu_2 - \lambda_2^2\mu_1}{\mu_1\mu_2(\lambda_1+\lambda_2)^2\lambda_2^2}(1 - e^{-\lambda_2 L\Delta}) - 2\frac{\lambda_1}{\lambda_2}\frac{Le^{-\lambda_2 L\Delta}\Delta}{\lambda_1+\lambda_2} \qquad (23) \end{aligned}$$

using that $E(A) = \kappa_1$ and $E(A^2) = \kappa_2$ are as given in Eqs. (16) and (17) above.

Finally, in case A_3 we have:

$$\min\{t_{i+U}^{*(1)}, t_i^{*(2)}\} - \max\{t_{i-L-1}^{*(1)}, t_{i-1}^{*(2)}\} = \min\{w_2, L\Delta + w_1\} - \overline{z}_1$$

where $\overline{z}_1 = z_1 - (U+1)\Delta$. In this case $\overline{z}_1$ and $C = \min\{w_2, L\Delta + w_1\}$ are not independent anymore because we impose the condition $0 < \overline{z}_1 < w_2$.

$$\begin{aligned} p(\overline{z}_1, w_1, w_2|\overline{z}_1 < L\Delta) &= \frac{\lambda_1 e^{-\lambda_1\overline{z}_1}\lambda_1 e^{-\lambda_1 w_1}\lambda_2 e^{-\lambda_2 w_2}}{1 - e^{-\lambda_1 L\Delta}} \\ p(\overline{z}_1, w_1, w_2|\overline{z}_1 < L\Delta \cap \overline{z}_1 < w_2) &= \frac{p(\overline{z}_1, w_1, w_2|\overline{z}_1 < L\Delta)}{\Pr\{\overline{z}_1 < w_2|\overline{z}_1 < L\Delta\}} \\ &= \frac{\lambda_1\lambda_2(\lambda_1+\lambda_2)e^{-\lambda_1\overline{z}_1 - \lambda_1 w_1 - \lambda_2 w_2}}{1 - e^{-L\Delta(\lambda_1+\lambda_2)}} \end{aligned}$$

using that

$$\begin{aligned} \Pr\{\overline{z}_1 < w_2|\overline{z}_1 < L\Delta\} &= \int_0^\infty\int_0^{L\Delta}\int_0^{w_2}\frac{\lambda_1 e^{-\lambda_1\overline{z}_1}\lambda_1 e^{-\lambda_1 w_1}\lambda_2 e^{-\lambda_2 w_2}}{1 - e^{-\lambda_1 L\Delta}}d\overline{z}_1 dw_2 dw_1 \\ &\quad + \int_0^\infty\int_{L\Delta}^\infty\int_0^{L\Delta}\frac{\lambda_1 e^{-\lambda_1\overline{z}_1}\lambda_1 e^{-\lambda_1 w_1}\lambda_2 e^{-\lambda_2 w_2}}{1 - e^{-\lambda_1 L\Delta}}d\overline{z}_1 dw_2 dw_1 \\ &= \frac{\lambda_1\left(1 - e^{-L\Delta(\lambda_1+\lambda_2)}\right)}{(\lambda_1+\lambda_2)(1 - e^{-\lambda_1 L\Delta})} \end{aligned}$$

With this we derive:

$$\begin{aligned} E(\overline{z}_1) &= \int_0^\infty\int_0^{L\Delta}\int_0^{w_2}\overline{z}_1\frac{\lambda_1\lambda_2(\lambda_1+\lambda_2)e^{-\lambda_1\overline{z}_1 - \lambda_1 w_1 - \lambda_2 w_2}}{1 - e^{-L\Delta(\lambda_1+\lambda_2)}}d\overline{z}_1 dw_2 dw_1 \\ &\quad + \int_0^\infty\int_{L\Delta}^\infty\int_0^{L\Delta}\overline{z}_1\frac{\lambda_1\lambda_2(\lambda_1+\lambda_2)e^{-\lambda_1\overline{z}_1 - \lambda_1 w_1 - \lambda_2 w_2}}{1 - e^{-L\Delta(\lambda_1+\lambda_2)}}d\overline{z}_1 dw_2 dw_1 \\ &= \frac{1 - (L\Delta(\lambda_1+\lambda_2)+1)e^{-L\Delta(\lambda_1+\lambda_2)}}{(\lambda_1+\lambda_2)\left(1 - e^{-L\Delta(\lambda_1+\lambda_2)}\right)} \end{aligned}$$

and similarly

$$E(\overline{z}_1^2) = \frac{2 - \left((L\Delta(\lambda_1+\lambda_2)+1)^2 + 1\right)e^{-L\Delta(\lambda_1+\lambda_2)}}{(\lambda_1+\lambda_2)^2\left(1 - e^{-L\Delta(\lambda_1+\lambda_2)}\right)}$$

The expectation of the minimum can be expressed as:

$$
\begin{aligned}
E\left(\min\left\{w_2, L\Delta + w_1\right\}|A_3\right) &= E\left(w_2|w_2 < L\Delta \cap A_3\right)\Pr\left\{w_2 < L\Delta|A_3\right\} \\
&\quad + \left(L\Delta + E\left(\min\left\{\overline{w}_2, w_1\right\}|A_3\right)\right)\Pr\left\{w_2 > L\Delta|A_3\right\} \\
&= \int_0^\infty \int_0^{L\Delta} \int_0^{w_2} w_2 \frac{\lambda_1\lambda_2\left(\lambda_1+\lambda_2\right)e^{-\lambda_1\overline{z}_1-\lambda_1 w_1-\lambda_2 w_2}}{1-e^{-L\Delta(\lambda_1+\lambda_2)}} d\overline{z}_1 dw_2 dw_1 \\
&\quad + \left(L\Delta + \frac{1}{\lambda_1+\lambda_2}\right)\frac{\left(\lambda_1+\lambda_2\right)\left(1-e^{-L\Delta\lambda_1}\right)e^{-L\Delta\lambda_2}}{\lambda_1\left(1-e^{-L\Delta(\lambda_1+\lambda_2)}\right)} \\
&= \frac{\lambda_1+2\lambda_2-\lambda_2\left(1+L\Delta\left(\lambda_1+\lambda_2\right)\right)e^{-L\Delta(\lambda_1+\lambda_2)}-\left(\lambda_1+\lambda_2\right)e^{-L\Delta\lambda_2}}{\lambda_2\left(\lambda_1+\lambda_2\right)\left(1-e^{-L\Delta(\lambda_1+\lambda_2)}\right)}
\end{aligned}
$$

and the squared minimum:

$$
\begin{aligned}
E(\min\left\{w_2, L\Delta + w_1\right\}^2|A_3) &= E\left(w_2^2|w_2 < L\Delta \cap A_3\right)\Pr\left\{w_2 < L\Delta|A_3\right\} \\
&\quad + \left(L\Delta + E\left(\min\left\{\overline{w}_2, w_1\right\}|A_3\right)\right)^2\Pr\left\{w_2 > L\Delta|A_3\right\} \\
&= \int_0^\infty \int_0^{L\Delta} \int_0^{w_2} w_2^2 \frac{\lambda_1\lambda_2\left(\lambda_1+\lambda_2\right)e^{-\lambda_1 z_1-\lambda_1 w_1-\lambda_2 w_2}}{1-e^{-L\Delta(\lambda_1+\lambda_2)}} dz_1 dw_2 dw_1 \\
&\quad + \left(L^2\Delta^2 + 2\frac{L\Delta}{\lambda_1+\lambda_2} + \frac{2}{\left(\lambda_1+\lambda_2\right)^2}\right)\frac{\left(\lambda_1+\lambda_2\right)\left(1-e^{-L\Delta\lambda_1}\right)e^{-L\Delta\lambda_2}}{\lambda_1\left(1-e^{-L\Delta(\lambda_1+\lambda_2)}\right)} \\
&= \frac{\begin{array}{c}\lambda_2^2\left(\left(L\Delta\left(\lambda_1+\lambda_2\right)+1\right)^2+1\right)e^{-L\Delta(\lambda_1+\lambda_2)}-2\lambda_1^2-6\lambda_1\lambda_2 \\ -6\lambda_2^2+2\left(\lambda_1+\lambda_2\right)\left(L\Delta\left(\lambda_1+\lambda_2\right)\lambda_2+2\lambda_2+\lambda_1\right)e^{-L\Delta\lambda_2}\end{array}}{\lambda_2^2\left(\lambda_1+\lambda_2\right)^2\left(-1+\exp\left(-L\Delta\left(\lambda_1+\lambda_2\right)\right)\right)}
\end{aligned}
$$

Finally, the cross product

$$
\begin{aligned}
E\left(\overline{z}_1\min\left\{w_2, L\Delta + w_1\right\}|A_3\right) &= E\left(\overline{z}_1 w_2|w_2 < L\Delta \cap A_3\right)\Pr\left\{w_2 < L\Delta|A_3\right\} \\
&\quad + \left(L\Delta + E\left(\min\left\{\overline{w}_2, w_1\right\}|A_3\right)\right)E\left(\overline{z}_1\right)\Pr\left\{w_2 > L\Delta|A_3\right\} \\
&= \frac{\lambda_1\lambda_2\left(\lambda_1+\lambda_2\right)}{1-e^{-L\Delta(\lambda_1+\lambda_2)}}\int_0^\infty \int_0^{L\Delta} \int_0^{w_2} z_1 w_2 e^{-\lambda_1 z_1-\lambda_1 w_1-\lambda_2 w_2} dz_1 dw_2 dw_1 \\
&\quad + \left(L\Delta + \frac{1}{\lambda_1+\lambda_2}\right)\frac{1-\left(\lambda_1 L\Delta+1\right)e^{-\lambda_1 L\Delta}}{\lambda_1\left(1-e^{-\lambda_1 L\Delta}\right)}\frac{\left(\lambda_1+\lambda_2\right)\left(1-e^{-L\Delta\lambda_1}\right)e^{-L\Delta\lambda_2}}{\lambda_1\left(1-e^{-L\Delta(\lambda_1+\lambda_2)}\right)} \\
&= \frac{\begin{array}{c}\lambda_1\left(3\lambda_2+\lambda_1\right)-\left(\lambda_1+\lambda_2\right)^2e^{-L\Delta\lambda_2} \\ -\lambda_2\left(\lambda_1-\lambda_2+\lambda_1 L\Delta\left(\lambda_1+\lambda_2\right)\left(L\Delta\left(\lambda_1+\lambda_2\right)+2\right)\right)e^{-L\Delta(\lambda_1+\lambda_2)}\end{array}}{\lambda_1\lambda_2\left(\lambda_1+\lambda_2\right)^2\left(1-e^{-L\Delta(\lambda_1+\lambda_2)}\right)}
\end{aligned}
$$

Combining all terms we get:

$$
E\left(\min\left\{w_2, L\Delta + w_1\right\}-\overline{z}_1\right)^2 = 2\frac{1-e^{-\lambda_2 L\Delta}-\lambda_2 L\Delta e^{-\lambda_2 L\Delta}}{\lambda_2^2\left(1-e^{-(\lambda_2+\lambda_1)L\Delta}\right)} + 2\frac{\left(1-e^{-\lambda_1 L\Delta}\right)e^{-\lambda_2 L\Delta}}{\lambda_1\left(\lambda_2+\lambda_1\right)\left(1-e^{-(\lambda_2+\lambda_1)L\Delta}\right)} \tag{24}
$$

Analogous to the discussion in section A.1.1, it can be shown (after tedious calculations omitted here) that $\sum_i\sum_{j\neq i}E(\overline{R}_i Z_i \overline{R}_j Z_j)$ is tightly bounded by $M(M-1)(E(\overline{R}_i Z_i))^2$ and so for simplicity of exposition we equate these terms. The variance expression in Eq. (9) with $\xi_1 = \xi_2 = 0$ then directly follows.

The noise contribution to the variance of RCLL is analogous to the RC case with the only difference that now returns on $\overline{R}_i$ are serially correlated (up to a displacement of $U+L$) due to lead-lag adjustment. In particular, we have:

$$\begin{aligned} E\left(\overline{R}_i\overline{R}_{i+h}\right) &= E\left(\sum_{j=-L}^{U} R_{i+j}\right)\left(\sum_{j=-L}^{U} R_{i+j+h}\right) = \sum_{j=\max\{-L,-L-h\}}^{\min\{U,U-h\}} E\left(R^2_{i+j+h}\right), \\ &= \sigma_1^2 \max\left\{0, U+L+1-|h|\right\}\Delta. \end{aligned}$$

Assuming, as before, that the contaminated returns are $\overline{R}_i + u_i - u_{i-1}$ and $Z_i + v_i - v_{i-1}$, then

$$\begin{aligned} V(RCLL_M^{(iid)}) - V(RCLL_M) &= \mu_1^*\mu_2 \sum_{i=1}^{M} E(2v_i^2\overline{R}_i^2 + 2u_i^2 Z_i^2 + 4u_i^2 v_i^2 | E_i^{(1)} \cap E_i^{(2)}) \\ &\quad +2(\mu_1^*\mu_1 + \mu_1^* - \mu_1)\mu_2^2 \sum_{i=1}^{M-1} E\left(u_i^2 v_i^2 | E_{i,i+1}^{(1)} \cap E_{i,i+1}^{(2)}\right) - 2\mu_2^2 \sum_{i=1}^{M-1} E\left(v_i^2 \overline{R}_i \overline{R}_{i+1} | E_{i,i+1}^{(2)}\right) \\ &= 2\mu_2\sigma_1^2\xi_2^2\left(U+L+1\right) + 2\mu_1^*\sigma_2^2\xi_1^2 + 4M\mu_1^*\mu_2\xi_1^2\xi_2^2 \\ &\quad +2\left(M-1\right)\left(\mu_1^*\mu_1 + \mu_1^* - \mu_1\right)\mu_2^2\xi_1^2\xi_2^2 - 2\left(M-1\right)\mu_2^2\xi_2^2\sigma_1^2\left(U+L\right)\Delta \end{aligned}$$

using that $E(v_i^2\overline{R}_i^2 | E_i^{(1)} \cap E_i^{(2)}) = \xi_2^2\sigma_1^2\left(U+L+1\right)\Delta/\mu_1^*$.

■

A.3 Proof of Theorem 2.3

To prove unbiasedness of HY simply note that:

$$E(HY) = \rho\sigma_1\sigma_2 \sum_{i=1}^{M_1}\sum_{j\in A_i} E(\nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)}))) = \rho\sigma_1\sigma_2$$

The variance can be expressed as:

$$\begin{aligned} V(HY) &= \sum_{i=1}^{M_1}\sum_{j\in A_i}\sum_{h=1}^{M_1}\sum_{l\in A_h} Cov(R_i Z_j, R_h Z_l) \\ &= \sum_{i=1}^{M_1}\sum_{j\in A_i}\sum_{h=1}^{M_1}\sum_{l\in A_h} Cov(R_i, R_h)Cov(Z_j, Z_l) + \sum_{i=1}^{M_1}\sum_{j\in A_i}\sum_{h=1}^{M_1}\sum_{l\in A_h} Cov(R_i, Z_l)Cov(Z_j, R_h) \\ &= \sum_{i=1}^{M_1}\sum_{j\in A_i} V(R_i)V(Z_j) + \sum_{i=1}^{M_1}\sum_{j\in A_i}\sum_{h=1}^{M_1}\sum_{l\in A_h} Cov(R_i, Z_l)Cov(Z_j, R_h) \\ &= \sigma_1^2\sigma_2^2 E\left(I_1\right) + \rho^2\sigma_1^2\sigma_2^2 E\left(I_2\right) \end{aligned}$$

where

$$\begin{aligned} I_1 &= \sum_{i=1}^{M_1}\sum_{j\in A_i} \nu(t_{i-1}^{(1)}, t_i^{(1)})\nu(t_{j-1}^{(2)}, t_j^{(2)}) \\ I_2 &= \sum_{i=1}^{M_1}\sum_{j\in A_i}\sum_{h=1}^{M_1}\sum_{l\in A_h} \nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{l-1}^{(2)}, t_l^{(2)}))\nu((t_{h-1}^{(1)}, t_h^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)})) \end{aligned}$$

The expectation of the first term, conditional on M_1, is equal to:

$$I_1 = \sum_{i=1}^{M_1} E(\nu(t_{i-1}^{(1)}, t_i^{(1)})(\nu(t_{i-1}^{(1)}, t_i^{(1)}) + x^\star)) = \left(\frac{2}{\lambda_1^2} + \frac{1}{\lambda_1}\frac{2}{\lambda_2}\right) M_1$$

where $x^\star = \sum_{j \in A_i} \nu(t_{j-1}^{(2)}, t_j^{(2)}) - \nu(t_{i-1}^{(1)}, t_i^{(1)})$. Throughout the proof we assume that the inter-arrival times are independent exponentially distributed random variables. Strictly speaking, when conditioning on M_1, the process is binomial but this distinction will be immaterial for typical values of λ_1 and λ_2 when taking expectations of functions of the inter arrival times. Next, taking expectations w.r.t. M_1 we have:

$$E(I_1) = 2\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)$$

The second term can be expressed as:

$$I_2 = \sum_{i=1}^{M_1} \left(\sum_{j \in A_i} \nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)})) \right)^2 + 2 \sum_{h=1}^{M_1} \sum_{j \in A_i} \sum_{i=1}^{h-1} \sum_{l \in A_h} \nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{l-1}^{(2)}, t_l^{(2)})) \nu((t_{h-1}^{(1)}, t_h^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)})).$$

Conditional on M_1, we have:

$$E \sum_{i=1}^{M_1} \left(\sum_{j \in A_i} \nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)})) \right)^2 = \sum_{i=1}^{M_1} E(\nu(t_{i-1}^{(1)}, t_i^{(1)})^2) = \frac{2M_1}{\lambda_1^2}$$

Taking expectations w.r.t. M_1, this term is equal to $2/\lambda_1$. Next, if $i < h$ the expectation of $\nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{l-1}^{(2)}, t_l^{(2)})) \nu((t_{h-1}^{(1)}, t_h^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)}))$ is non-zero only if asset 2 does not transact on the interval $[t_i^{(1)}, t_{h-1}^{(1)}]$. Conditional on $t^{(1)}$ we then have:

$$\begin{aligned}
& E\left(\nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{l-1}^{(2)}, t_l^{(2)})) \nu((t_{h-1}^{(1)}, t_h^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)}))\right) \\
= \; & \exp\{-\lambda_2(t_{h-1}^{(1)} - t_i^{(1)})\} \left[\exp\{-\lambda_2(t_i^{(1)} - t_{i-1}^{(1)})\}(t_i^{(1)} - t_{i-1}^{(1)}) + \frac{1 - \left(\lambda_2(t_i^{(1)} - t_{i-1}^{(1)}) + 1\right) e^{-\lambda_2(t_i^{(1)} - t_{i-1}^{(1)})}}{\lambda_2} \right] \\
& \times \left[\exp\{-\lambda_2(t_h^{(1)} - t_{h-1}^{(1)})\}(t_h^{(1)} - t_{h-1}^{(1)}) + \frac{1 - \left(\lambda_2(t_h^{(1)} - t_{h-1}^{(1)}) + 1\right) e^{-\lambda_2(t_h^{(1)} - t_{h-1}^{(1)})}}{\lambda_2} \right]
\end{aligned}$$

Using that $t_i^{(1)} - t_{i-1}^{(1)}$ is exponentially distributed with parameter λ_1 and $t_{h-1}^{(1)} - t_i^{(1)}$ is gamma distributed with parameters $(h - i - 1, \lambda_1)$, we get:

$$\begin{aligned}
E\left(\exp\{-\lambda_2(t_{h-1}^{(1)} - t_i^{(1)})\}\right) &= \left(\frac{\lambda_1}{\lambda_2 + \lambda_1}\right)^{h-i-1} \\
E\left((t_i^{(1)} - t_{i-1}^{(1)}) \exp\{-\lambda_2(t_i^{(1)} - t_{i-1}^{(1)})\}\right) &= \frac{\lambda_1}{(\lambda_2 + \lambda_1)^2}
\end{aligned}$$

Using this we obtain:

$$E\left(\nu((t_{i-1}^{(1)}, t_i^{(1)}) \cap (t_{l-1}^{(2)}, t_l^{(2)})) \nu((t_{h-1}^{(1)}, t_h^{(1)}) \cap (t_{j-1}^{(2)}, t_j^{(2)}))\right) = \frac{1}{(\lambda_2 + \lambda_1)^2} \left(\frac{\lambda_1}{\lambda_2 + \lambda_1}\right)^{h-i-1}$$

Summing and taking expectations w.r.t. M_1 we get:

$$\begin{aligned}
E \frac{1}{(\lambda_2 + \lambda_1)^2} \sum_{h=1}^{M_1} \sum_{i=1}^{h-1} \left(\frac{\lambda_1}{\lambda_2 + \lambda_1}\right)^{h-i-1} &= \frac{1}{(\lambda_2 + \lambda_1)^2} E\left((\lambda_1 + \lambda_2) \frac{\lambda_1}{\lambda_2} - \frac{(\lambda_1 + \lambda_2)^2}{\lambda_2^2} + \frac{(\lambda_1 + \lambda_2)^2}{\lambda_2^2} \left(\frac{\lambda_1}{\lambda_2 + \lambda_1}\right)^{M_1} \right) \\
&= \frac{1}{\lambda_1 + \lambda_2} \frac{\lambda_1}{\lambda_2} + \frac{1}{\lambda_2^2} \left(e^{-\frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}} - 1 \right)
\end{aligned}$$

The leading term in this expression is the first. For reasonable values of λ, the second term is negligible and we suspect it is accounting for end-effects so we ignore it. Collecting terms we then get the required expression in Eq. (13) for the no-noise case.

With i.i.d. noise, $V(HY) = E(HY^2) - E(HY)^2$, where $E(HY)$ is as before.

$$E(HY^2) = \sum_{i=1}^{M_1} \sum_{j \in A_i} \sum_{h=1}^{M_1} \sum_{k \in A_h} E((R_i + u_i - u_{i-1})(Z_j + v_j - v_{j-1})(R_h + u_h - u_{h-1})(Z_k + v_k - v_{k-1}))$$

If $i = h$ then

$$\begin{aligned}
&\sum_{j \in A_i} \sum_{k \in A_i} E((R_i + u_i - u_{i-1})^2 (Z_j + v_j - v_{j-1})(Z_k + v_k - v_{k-1})) \\
&= \sum_{j \in A_i} \sum_{k \in A_i} E(R_i^2 Z_j Z_k) + \sum_{j \in A_i} \sum_{k \in A_i} E(R_i^2) E((v_j - v_{j-1})(v_k - v_{k-1})) \\
&\quad + \sum_{j \in A_i} \sum_{k \in A_i} E((u_i - u_{i-1})^2) E(Z_j Z_k) + \sum_{j \in A_i} \sum_{k \in A_i} E((u_i - u_{i-1})^2) E((v_j - v_{j-1})(v_k - v_{k-1}))
\end{aligned}$$

The first-order MA structure of $v_j - v_{j-1}$ implies that

$$\sum_{j \in A_i} \sum_{k \in A_i} E((v_j - v_{j-1})(v_k - v_{k-1})) = (2\#A_i - 2(\#A_i - 2) - 2)\xi_2^2 = 2\xi_2^2$$

where $\#B$ denotes the number of distinct elements in the set B, so that

$$\begin{aligned}
&\sum_{j \in A_i} \sum_{k \in A_i} E((R_i + u_i - u_{i-1})^2 (Z_j + v_j - v_{j-1})(Z_k + v_k - v_{k-1})) \\
&= \sum_{j \in A_i} \sum_{k \in A_i} E(R_i^2 Z_j Z_k) + 2\xi_2^2 E(R_i^2) + 2\xi_1^2 \sum_{j \in A_i} E(Z_j^2) + 4\xi_1^2 \xi_2^2
\end{aligned}$$

and it follows that

$$\begin{aligned}
&\sum_{i=1}^{M_1} \sum_{j \in A_i} \sum_{k \in A_i} E((R_i + u_i - u_{i-1})^2 (Z_j + v_j - v_{j-1})(Z_k + v_k - v_{k-1})) \\
&= \sum_{i=1}^{M_1} \sum_{j \in A_i} \sum_{k \in A_i} E(R_i^2 Z_j Z_k) + 2\xi_2^2 \sigma_1^2 + 2\xi_1^2 \sigma_2^2 \left(1 + \frac{2\lambda_1}{\lambda_2}\right) + 4\lambda_1 \xi_1^2 \xi_2^2
\end{aligned}$$

If $i \neq h$

$$\begin{aligned}
&\sum_{j \in A_i} \sum_{k \in A_h} E((R_i + u_i - u_{i-1})(Z_j + v_j - v_{j-1})(R_h + u_h - u_{h-1})(Z_k + v_k - v_{k-1})) \\
&= \sum_{j \in A_i} \sum_{k \in A_h} E(R_i Z_j R_h Z_k) + E((u_h - u_{h-1})(u_i - u_{i-1})) \sum_{j \in A_i} \sum_{k \in A_h} E(Z_j Z_k) \\
&\quad + E((u_i - u_{i-1})(u_h - u_{h-1})) \sum_{j \in A_i} \sum_{k \in A_h} E((v_j - v_{j-1})(v_k - v_{k-1})) \\
&= \sum_{j \in A_i} \sum_{k \in A_h} E(R_i Z_j R_h Z_k) - (I(h = i+1) + I(h = i-1))\xi_1^2 \sum_{j \in A_i} \sum_{k \in A_h} E(Z_j Z_k) \\
&\quad - (I(h = i+1) + I(h = i-1))\xi_1^2 \sum_{j \in A_i} \sum_{k \in A_h} E((v_j - v_{j-1})(v_k - v_{k-1}))
\end{aligned}$$

The MA structure of $v_j - v_{j-1}$ implies that $E((v_j - v_{j-1})(v_k - v_{k-1}))$ will be non-zero only if $k = j-1$, $k = j$, $k = j+1$. For $j \in A_i$ and $k \in A_{i+1}$, then there is only one for which $k = j = x$, and if $\#A_{i+1} > 1$ then $x + 1$ must also be in A_{i+1}.

$$\sum_{j\in A_i}\sum_{k\in A_h} E((v_j - v_{j-1})(v_k - v_{k-1})) = \xi_2^2(I(h = i+1) + I(h = i-1))(2 - I(\#A_i > 1) - I(\#A_h > 1))$$

$$\sum_{j\in A_i}\sum_{k\in A_h} E(Z_j Z_k) = \frac{1}{\lambda_2}$$

It is easy to show that $E[I(\#A_i) > 1)] = \frac{\lambda_2}{\lambda_1+\lambda_2}$ and so

$$\begin{aligned}
&\sum_{j\in A_i}\sum_{k\in A_h} E((R_i + u_i - u_{i-1})(Z_j + v_j - v_{j-1})(R_h + u_h - u_{h-1})(Z_k + v_k - v_{k-1})) \\
=&\sum_{j\in A_i}\sum_{k\in A_h} E(R_i Z_j R_h Z_k) - (I(h = i+1) + I(h = i-1))\xi_1^2\frac{1}{\lambda_2} \\
&- 2(I(h = i+1) + I(h = i-1))\xi_1^2\xi_2^2\frac{\lambda_1}{\lambda_1+\lambda_2}
\end{aligned}$$

Thus:

$$\begin{aligned}
&\sum_{i=1}^{M_1}\sum_{h=1}^{M_1}\sum_{j\in A_i}\sum_{k\in A_h} E((R_i + u_i - u_{i-1})(Z_j + v_j - v_{j-1})(R_h + u_h - u_{h-1})(Z_k + v_k - v_{k-1})) \\
=&\sum_{i=1}^{M_1}\sum_{j\in A_i}\sum_{h=1}^{M_1}\sum_{k\in A_h} E(R_i Z_j R_h Z_k) - \xi_1^2\sigma_1^2 2(\lambda_1 - 1)\frac{1}{\lambda_2} - 4\xi_1^2\xi_2^2(\lambda_1 - 1)\frac{\lambda_1}{\lambda_1+\lambda_2}
\end{aligned}$$

and

$$V(HY) = V(HY^{\text{no noise}}) + 2\xi_2^2\sigma_1^2 + 2\xi_1^2\sigma_2^2 + 4\xi_1^2\xi_2^2\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}$$

■

Figure 7: Empirical and model-implied RC signature plots - NYSE quotes

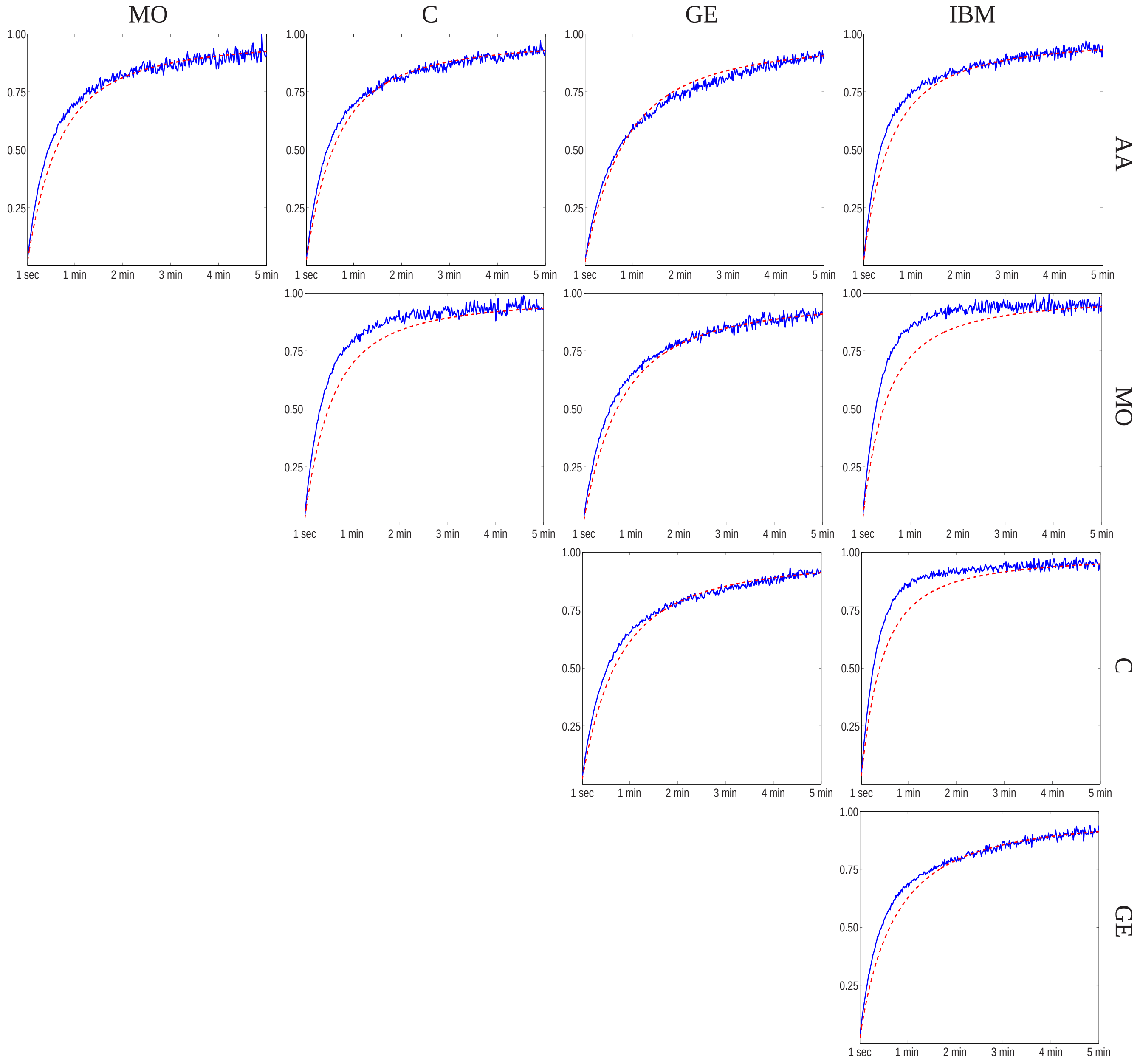


Figure 8: Empirical and model-implied RCLL(1,1) signature plots - NYSE quotes

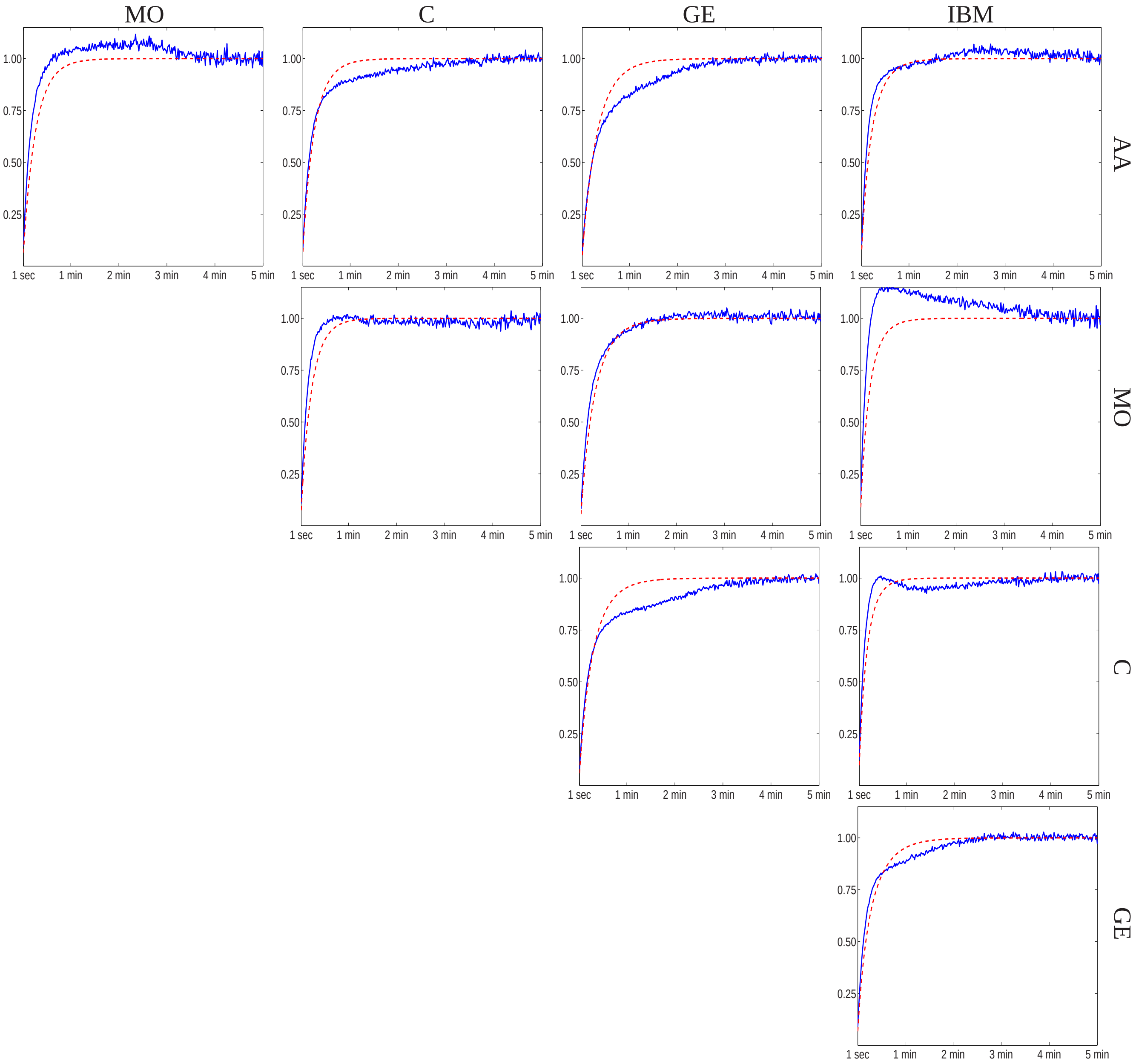


Figure 9: Empirical HY covariance signature plots - NYSE quotes

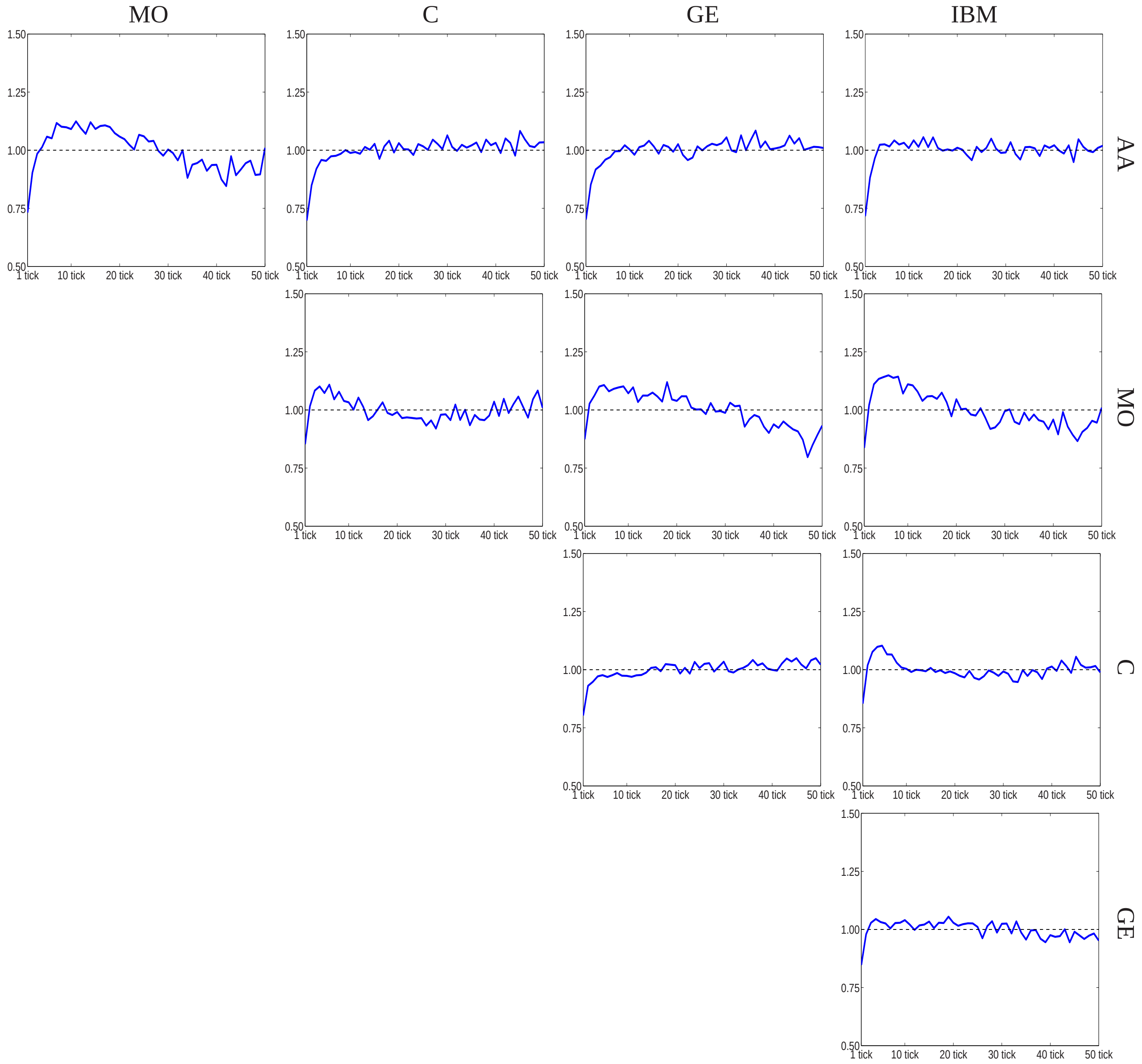


Figure 10: Empirical and model-implied RC signature plots - NYSE transactions

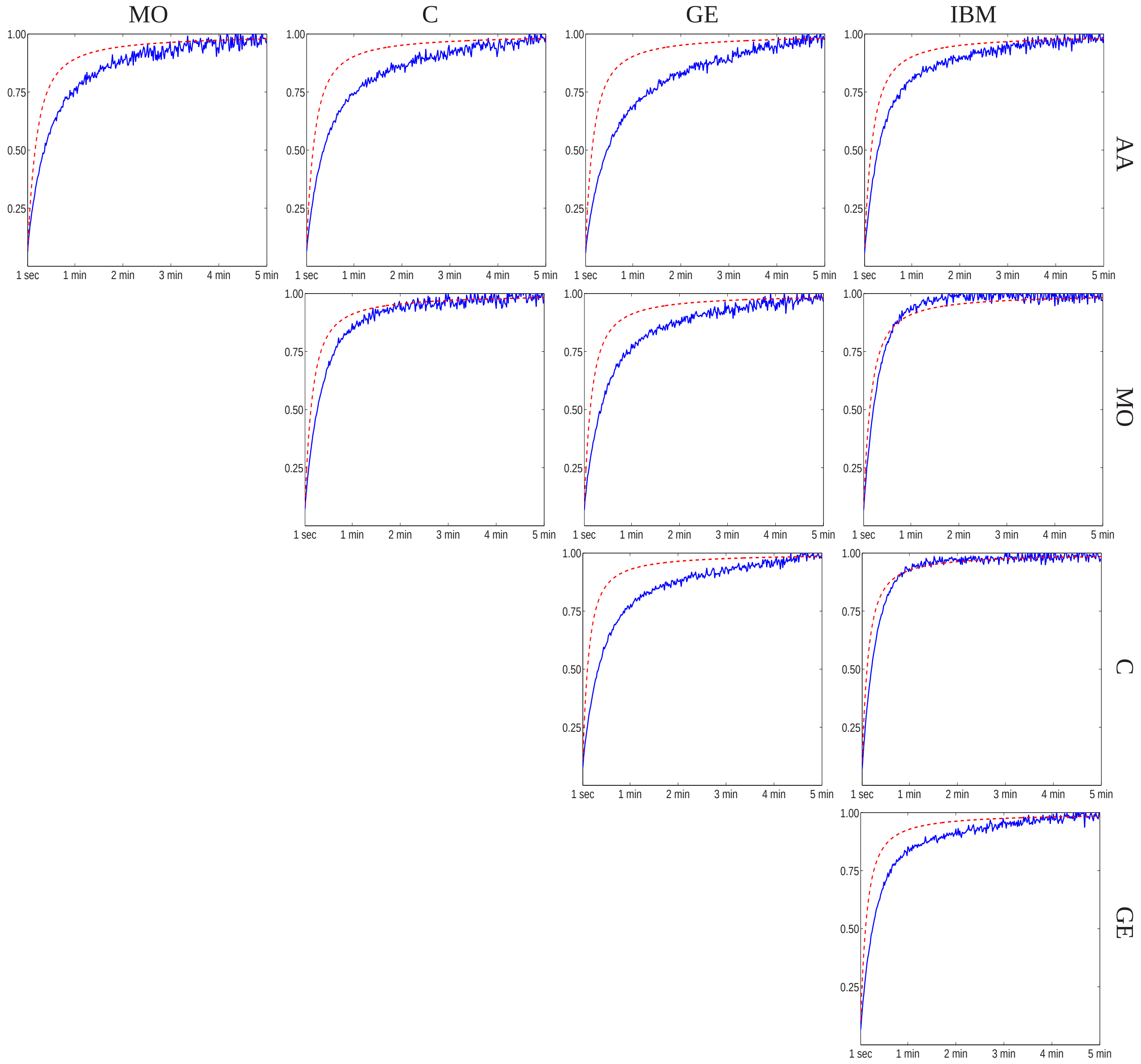


Figure 11: Empirical and model-implied RCLL(1,1) signature plots - NYSE transactions

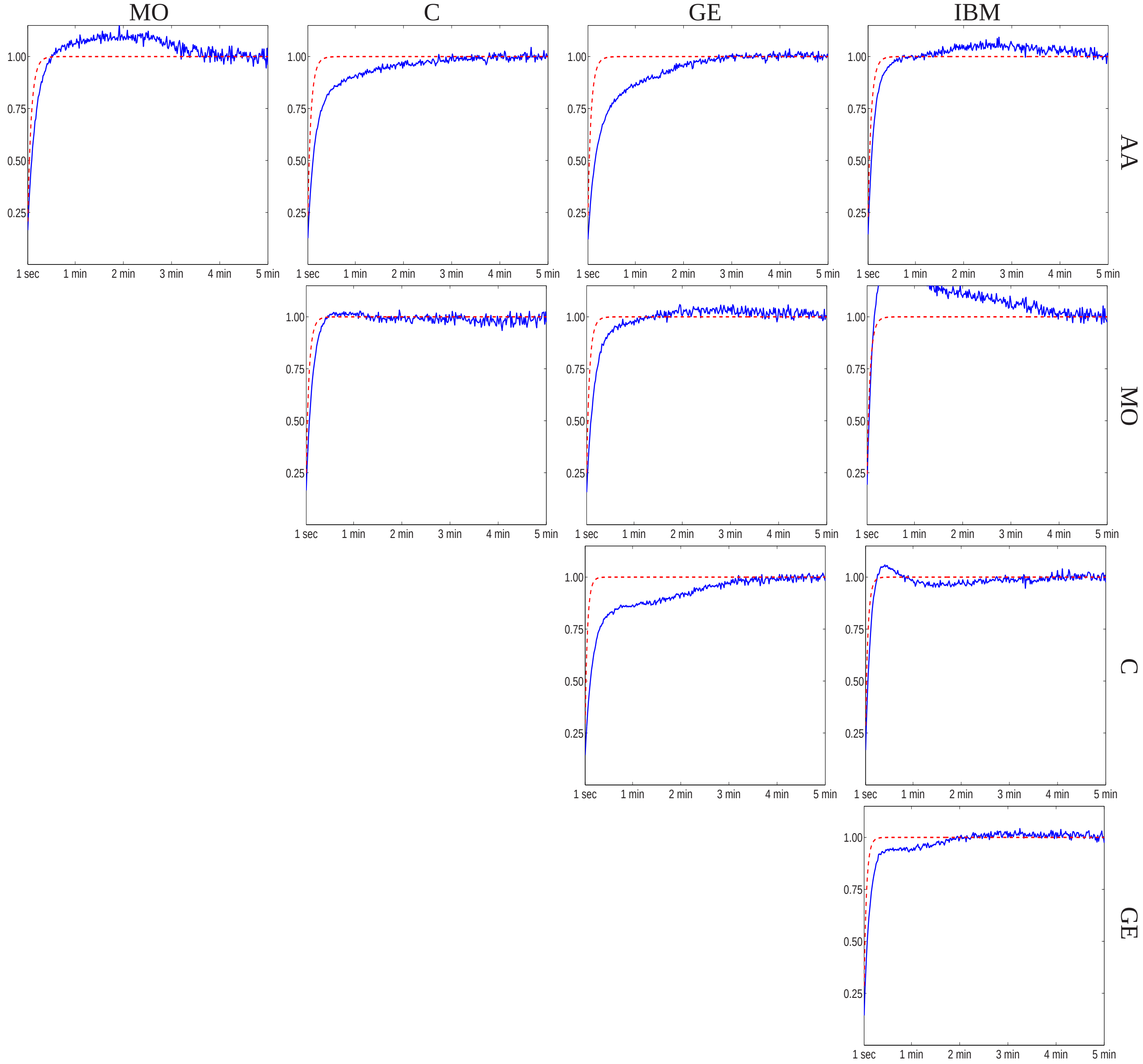
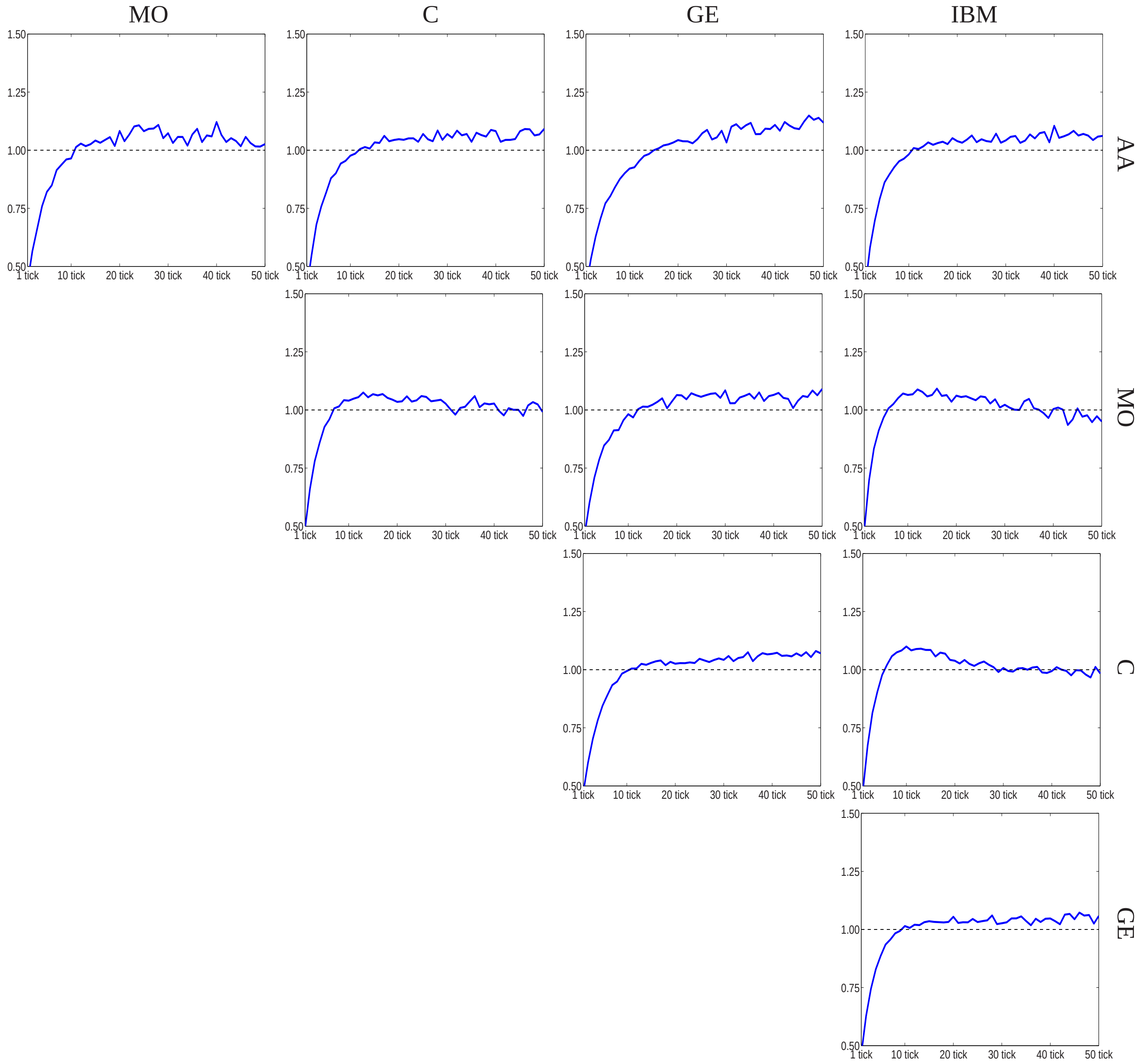


Figure 12: Empirical HY covariance signature plots - NYSE transactions



References

Aït-Sahalia, Y., P. Mykland, and L. Zhang, 2005, “How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise,” *Review of Financial Studies*, 18 (2), 351–416.

Andersen, T. G., and T. Bollerslev, 1998, “Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts,” *International Economic Review*, 39 (4), 885–905.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys, 2003, “Modeling and Forecasting Realized Volatility,” *Econometrica*, 71 (2), 579–625.

Atchison, M. D., K. C. Butler, and R. R. Simonds, 1987, “Nonsynchronous Security Trading and Market Index Autocorrelation,” *Journal of Finance*, 42 (1), 111–118.

Bandi, F. M., and J. R. Russell, 2005, “Realized Covariation, Realized Beta, and Microstructure Noise,” manuscript University of Chicago.

——— , 2006, “Separating Microstructure Noise from Volatility,” *Journal of Financial Economics*, 79, 655–692.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard, 2006, “Designing Realised Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise,” manuscript University of Oxford, Nuffield College.

Barndorff-Nielsen, O. E., and N. Shephard, 2004, “Econometric Analysis of Realised Covariation: High Frequency Based Covariance, Regression and Correlation in Financial Economics,” *Econometrica*, 72 (3), 885–925.

Bollerslev, T., and B. Y. Zhang, 2003, “Measuring and Modeling Systematic Risk in Factor Pricing Model Using High-Frequency Data,” *Journal of Empirical Finance*, 10, 533–558.

Chordia, T., and B. Swaminathan, 2000, “Trading Volume and Cross-Autocorrelations in Stock Returns,” *Journal of Finance*, 55 (2), 913–935.

Cohen, K. J., G. A. Hawawini, S. F. Maier, R. A. Schwartz, and D. K. Whitcomb, 1983, "Friction in the Trading Process and the Estimation of Systematic Risk," *Journal of Financial Economics*, 12, 263–278.

Corsi, F., 2006, "Realized Correlation Tick-by-Tick," manuscript University of Lugano, Institute of Finance.

Corsi, F., G. Zumbach, U. A. Müller, and M. M. Dacorogna, 2001, "Consistent High-Precision Volatility from High-Frequency Data," *Economic Notes*, 30 (2), 183–204.

de Jong, F., and T. Nijman, 1997, "High Frequency Analysis of Lead-Lag Relationships Between Financial Markets," *Journal of Empirical Finance*, 4, 257–277.

Dimson, E., 1979, "Risk Measurement When Shares are Subject to Infrequent Trading," *Journal of Financial Economics*, 7, 197–226.

Epps, T. W., 1979, "Comovements in Stock Prices in the Very Short Run," *Journal of the American Statistical Association*, 74 (366), 291–298.

Fisher, L., 1966, "Some New Stock-Market Indexes," *Journal of Business*, 39 (1-2), 191–225.

Griffin, J. E., and R. C. Oomen, 2005, "Sampling Returns for Realized Variance Calculations: Tick Time or Transaction Time?," forthcoming *Econometric Reviews*.

Hansen, P. R., and A. Lunde, 2006, "Realized Variance and Market Microstructure Noise," *Journal of Business & Economic Statistics*, 24 (2), forthcoming.

Hayashi, T., and S. Kusuoka, 2004, "Nonsynchronous Covariation Measurement for Continuous Semimartingales," manuscript University of Tokyo, Graduate School of Mathematical Sciences.

Hayashi, T., and N. Yoshida, 2005, "On Covariance Estimation of Non-Synchronously Observed Diffusion Processes," *Bernoulli*, 11 (2), 359–379.

——— , 2006, "Estimating Correlations with Nonsynchronous Observations in Continuous Diffusion Models," manuscript Columbia University, Department of Statistics.

Lo, A. W., and A. C. MacKinlay, 1990, “An Econometric Analysis of Nonsynchronous-Trading,” *Journal of Econometrics*, 45, 181–212.

Martens, M., 2004, “Estimating Unbiased and Precise Realized Covariances,” manuscript Erasmus University Rotterdam.

Niederhoffer, V., and M. F. M. Osborne, 1966, “Market Making and Reversal on the Stock Exchange,” *Journal of the American Statistical Association*, 61 (316), 897–916.

Oomen, R. C., 2005, “Properties of Bias-Corrected Realized Variance under Alternative Sampling Schemes,” *Journal of Financial Econometrics*, 3 (4), 555–577.

——— , 2006, “Properties of Realized Variance under Alternative Sampling Schemes,” *Journal of Business & Economic Statistics*, 24 (2), 219 – 237.

Scholes, M., and J. Williams, 1977, “Estimating Betas from Nonsynchronous Data,” *Journal of Financial Economics*, 5, 309–327.

Shanken, J., 1987, “Nonsynchronous Data and the Covariance-Factor Structure of Returns,” *Journal of Finance*, 42 (2), 221–231.

Sheppard, K., 2005, “Realized Covariance and Scrambling,” manuscript University of Oxford.

Voev, V., and A. Lunde, 2005, “Integrated Covariance Estimation Using High-Frequency Data in the Presence of Noise,” manuscript University of Constance.

Zhang, L., 2006, “Estimating Covariation: Epps Effect, Microstructure Noise,” manuscript Carnegie Mellon University, Department of Statistics.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia, 2005, “A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data,” *Journal of the American Statistical Association*, 100, 1394–1411.

Zhou, B., 1996, “High Frequency Data and Volatility in Foreign-Exchange Rates,” *Journal of Business & Economic Statistics*, 14(1), 45–52.