Obtaining and Predicting the Bounds of Realized Correlations

Lidan Grossmass[a]

JEL-Classification: C14, C18, C58, G17
Keywords: High Frequency Data, Realized Covariance, Partial Identification, Bounds

1. Introduction

Given the plethora of bias correction methods for the estimation of realized covariance and correlation that only work well under certain conditions, this paper proposes a different approach to the problem. We argue that the inherent data problems render point identification of realized covariance and correlation unreliable, especially when the level of asynchronicity and microstructure noise is high. Under such circumstances, the data only allows for partial identification (Manski, 1995) of the realized covariance and correlation, whereas point identification of these measures requires prior assumptions about the data. Given the data limitations, partial identification analysis identifies the bounds that the mean of the distribution of interest lies in. Although conservative, the estimation of bounds is a more robust approach when estimating realized correlations, because both econometricians and practitioners should be aware of the worst and best case scenarios when assumptions about the data conditions that are needed to yield point estimations cannot or should not be made.

The availability of high frequency data in recent years has allowed financial econometrics to shift away from parametric conditional variance and covariance estimation based on daily or weekly data towards nonparametric ex-post measures termed as realized measures. Barndorff-Nielsen and Shephard (2002)

a University of Konstanz, Department of Economics, Box 124, 78457 Konstanz, Germany; Email: rachellidan@gmail.com; The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7-PEOPLE-ITN-2008 under grant agreement number PITN-GA-2009-237984. The funding is gratefully acknowledged. We would also like to thank Klaus Neusser (the Editor), an anonymous referee, Charles Manski, Winfried Pohlmeier, Peter Reinhard Hansen and Hao Liu for their insightful comments and suggestions.

and MYKLAND and ZHANG (2006) have shown that as sampling intervals get smaller, the realized variance or covariance is a consistent estimator of integrated variation or covariation and has an asymptotic variance that is mixture normally distributed. In this paper we consider the realized correlation, which is analogous to the realized covariance. It is almost trivial to state that the estimation of the correlation structure of asset returns is important for many areas in finance, such as in risk management and portfolio optimization.

The estimation of realized covariance (*RC*) encounters several problems. First of all, trading of different assets rarely occur simultaneously, i.e. trading is asynchronous. This causes the realized covariance to tend to zero as sampling frequency increases, a phenomenon termed as "Epps effect" (EPPS, 1979). The most common approach to estimating realized covariance is to construct approximately synchronised pairs using either previous-tick interpolation or linear interpolation. However asynchronicity renders these interpolated estimators to be biased (see e.g. DACOROGNA et al., 2001) and ZHANG (2011) derives the analytical bias of the previous-tick *RC*.

To correct for the bias due to asynchronicity, BARNDORFF-NIELSEN et al. (2011) propose the incorporation of lead and lag autocovariance terms based on the idea that returns sampled at regular calendar time will correlate with preceding and succeeding returns on other assets even if the underlying correlation is purely contemporaneous. Another approach to dealing with asynchronicity is the Hayashi and Yoshida estimator (HAYASHI and YOSHIDA, 2005) which accumulates cross-products of all fully and partially overlapping event-time returns to obtain unbiased covariance estimators.

Unfortunately asynchronicity is not the sole problem that realized covariance and correlation encounter. A further problem is that high frequency data is characteristically plagued by what is termed as "market microstructure noise", a distortion of the true latent efficient (or "frictionless") price which according to classical market microstructure theory (O'HARA, 1995) should evolve as a martingale. The existence of market microstructure noise is explained by market frictions that distort efficient prices (ROLL, 1984). These frictions could be induced by price discreteness, the existence of bid-ask spreads or the lack of liquidity. Market microstructure noise renders the true price process unobservable. AÏT-SAHALIA, MYKLAND, and ZHANG (2005) and ZHANG, MYKLAND, and AÏT-SAHALIA (2005) showed that the realized variance estimator (the realized covariance in the univariate case) is biased in the presence of market microstructure noise, and the bias becomes larger as the sampling frequency increases. This led to the literature about obtaining the optimal sampling frequency to reduce the effects of noise in a bias-variance tradeoff (see e.g. ZHOU, 1996; BANDI and RUSSELL, 2008).

To deal with the combination of both problems, methods such as subsampling (Zhang, Mykland, and Aït-Sahalia, 2005), pre-averaging (Jacod et al., 2009) and the two- and multi-scales estimators (Zhang, 2011) have been proposed to restore consistency of the estimators. Voev and Lunde (2007) proposed a bias correction method for the Hayashi and Yoshida estimator in the presence of dependent microstructure noise while Nolte and Voev (2009) propose a least squares approach to obtain the unbiased integrated volatility or co-volatility. Griffin and Oomen (2011) however showed that under a high enough noise level and low degree of correlation, the previous-tick *RC* that is not bias-corrected may be more efficient in terms of log mean-squared error than the Hayashi and Yoshida estimator and the lead-lag estimator of Barndorff-Nielsen et al. (2011).

Whichever the bias-correction method of the realized covariance and correlation, they require making some assumptions about the data problems and these assumptions may or may not be fulfilled in practice. We take instead a conservative approach by seeking identification bounds in the spirit of Manski (1995) on the previous-tick covariance estimator and the corresponding realized correlation. The issue of asynchronicity is regarded as a missing data problem because in the case of the previous-tick estimator, the problem is that the true latent price of one or more assets is not observed at the sampling time. We use the partial identification approach of Horowitz and Manski (2006) for incomplete data to obtain bounds of the identification region. The presence of microstructure noise can be regarded as a data corruption or contamination problem because the issue is that the sampling process is a distortion of the true latent price process. We use the approach by Horowitz and Manski (1995) for the treatment of contaminated and corrupted data to estimate the identification region.

The estimated bounds provide the worst and best case scenarios that can be found using information that the data provides without having to make assumptions about the inherent data problems. They require no structure to be imposed on the sample space and are attempts to guard against the worst outcomes that the data problems could possibly produce by using ex-post knowledge of the data. Altogether this serves as a more robust approach to inference, which is especially valuable when the realized covariance and correlation are used for estimating other useful measures such as sharpe ratios or the Value-at-Risk.

The paper is organised as follows: Section 2 gives the mathematical description of the previous tick realized covariance and correlation as well as the subsampled estimator of Zhang, Mykland, and Aït-Sahalia (2005); Section 3 describes the idea of partial identification, how we apply such identification analysis to estimate bounds of the realized covariance and correlation when the problems

of asynchronicity and microstructure noise are present in the data, some practical issues involved in the bounds estimation, and the forecasting of the bounds; Section 4 gives the results of a simulation exercise to study the efficacy of these bounds and their sensitivity to the tuning parameters; Section 5 gives an empirical application using two stocks, first describing the dataset and then the results of the bounds estimation and forecasting efficacy. Finally Section 6 concludes.

2. Realized Covariance and Correlation

Consider the price processes of two assets $\{X_t\}$ and $\{Y_t\}$. To estimate their integrated covariation $\langle X,Y\rangle_T$, the standard assumption is that both processes follow an Itô stochastic process with standard Brownian motion B^X and B^Y. Further assume that the processes have a drift coefficient μ_t^X and μ_t^Y with instantaneous variance $\sigma_t^{2,X}$ and $\sigma_t^{2,Y}$. Under such assumptions, the integrated covariation is given by

$$\langle X,Y\rangle_T = \int_0^T \sigma_t^X \sigma_t^Y d\langle B^X, B^Y\rangle_t. \tag{1}$$

BARNDORFF-NIELSEN and SHEPHARD (2002) and MYKLAND and ZHANG (2006) show that using the limit theorem for stochastic processes, an estimator for the integrated covariation is the realized covariance,

$$RC_T = \sum_{i:\tau_i\in[0,T]} (X_{\tau_i} - X_{\tau_{i-1}})(Y_{\tau_i} - Y_{\tau_{i-1}}), \tag{2}$$

which is a consistent estimator as sampling intervals get smaller and has an asymptotic mixture normal distribution. The realized correlation *RCorr* is obtained by dividing realized covariance by the realized volatilities (square root of realized variances) of the individual assets:

$$RCorr_T = \frac{RC_T}{\sqrt{\sum_{i:\tau_i\in[0,T]} (X_{\tau_i} - X_{\tau_{i-1}})^2}\sqrt{\sum_{i:\tau_i\in[0,T]} (Y_{\tau_i} - Y_{\tau_{i-1}})^2}}. \tag{3}$$

RC is most commonly estimated using previous-tick interpolation (or last-tick interpolation) which was found to be less biased than linear interpolation (Dacorogna et al 2001). As the name implies, the previous-tick *RC* is simply *RC*

estimated using prices on or immediately preceding a regularly spaced sampling time grid.

Consider a fixed time period $[0,T]$, usually a single trading day, to have a regularly sampled time grid denoted by $\mathcal{T}_N \in [0,\mathcal{T}]$, $\mathcal{T}_N = \{\tau_0,\tau_1,\ldots,\tau_{M_N}\}$, where M_N is the sampling frequency, N is the total number of observations or transactions of both X and Y, and $\Delta t = \tau_i - \tau_{i-1}$, $\forall i$, is the regular sampling interval that is constant, for example 1, 5, or 10 minutes. Hence a 5 minutes *RC* refers to *RC* with $\Delta t = 300$ seconds. We can view M_N as a filter on N and it is a function of Δt.

Let the irregular transaction times of X and Y be denoted by grids $\mathcal{G}_A$ and $\mathcal{H}_B$ respectively, with $\mathcal{G}_A = \{g_0, g_1,\ldots,g_A\}$ and $\mathcal{H}_B = \{h_0, h_1,\ldots,h_B\}$. Hence $N = A + B$. Also, $0 \leq g_0 \leq g_1 \leq \ldots \leq g_A \leq T$ and $0 \leq h_0 \leq h_1 \leq \ldots \leq h_B \leq T$. The previous ticks are then defined to be $a_i = \max\{g \in \mathcal{G}_A : g \leq \tau_i\}$ and $b_i = \max\{h \in \mathcal{H}_B : h \leq \tau_i\}$. The previous tick *RC* is thus given by

$$[X,Y]_T = \sum_{i=1}^{M_N} (X_{a_i} - X_{a_{i-1}})(Y_{b_i} - Y_{b_{i-1}}). \tag{4}$$

The bias of the previous tick *RC* due to asynchronicity alone (assuming no microstructure noise) is

$$-\int_0^T \langle X,Y \rangle'_u dF_N(u) + O_p\left(\frac{1}{N}\right) \text{ where } F_N(t) = \sum\nolimits_{i:\max(a_i,b_i)\leq t} |a_i - b_i|$$

(Zhang, 2011[1]), under the assumption that there is at least one pair of observations (g,h) within each $[\tau_i, \tau_{i+1}]$. While this condition is usually satisfied for highly traded assets, it does not hold for less liquid assets or in times of sudden liquidity shortages ('liquidity black holes') and inconsistency of the bias-corrected previous tick RC estimator results. Furthermore, the standard approach to characterise the bias due to asynchronicity is to assume a Poisson arrival rate for an observation (i.e. a trade or quote) that is independent of the price process (see for example Assumption 2 in Griffin and Oomen, 2011, and Section 6 of Zhang, 2011). Renault and Werker (2011) however showed that durations and price processes are not independent but exhibit instantaneous causality. This suggests

1 See Zhang (2011) which gives the analytical characterisation of the previous tick *RC* estimator in the presence of asynchronous trading and microstructure noise. Analytical solution of the bias can be obtained by assuming that the transaction time arrival rates follow independent intensity processes.

that the estimated bias due to asynchronicity would be inaccurate. The correction for the effect of microstructure noise involves empirical and theoretical modelling subtleties, because both microstructure noise and the efficient price are latent variables and a direct measurement of microstructure noise is not possible. ZHANG (2011) derives the bias of the previous-tick RC due to microstructure noise by assuming additive noise processes, with the analytical solution being available when the noise processes of X and Y (ε^X and ε^Y processes) are assumed to be independent white noise. However PHILLIPS and YU (2006) argue that the complexity of microstructure noise cannot be adequately captured by a simple white noise specification. They show that the properties of microstructure noise evolve over time, and may exhibit local non-stationarity and perfect correlation with the efficient price.

ZHANG, MYKLAND, and AÏT-SAHALIA (2005) proposed subsampling the estimator which reduces the bias caused by asynchronicity and microstructure noise. The subsampled estimator $[X,Y]_T^{(avg)}$ is constructed by averaging the estimators $[X,Y]_T^{(k)}$ across K grids, where the original regular spaced grid $\mathcal{T}$ is partitioned into K non-overlapping subgrids $\mathcal{T}^{(k)}$, $k=1,\ldots,K$. Consider the original grid $\mathcal{T}=\{\tau_0,\tau_1,\ldots,\tau_{M_N}\}$, then the k^{th} subgrid is given by $\mathcal{T}^{(k)}=\{\tau_{k-1},\tau_{k-1+K},\ldots,\tau_{k-1+M_NK}\}$. This creates K realized covariance estimates $[X,Y]_T^{(k)}, k=1,\ldots,K$ and the subsampled *RC* (*ssRC*) estimator is obtained by averaging:

$$ssRC_T=[X,Y]_T^{(avg)}=\frac{1}{K}\sum_{k=1}^{K}[X,Y]_T^{(k)} \tag{5}$$

and the corresponding subsampled *RCorr* (*ssRCorr*) estimator is

$$ssRCorr_T=[X,Y]_T^{(avg)}=\frac{1}{K}\sum_{k=1}^{K}\frac{[X,Y]_T^{(k)}}{\sqrt{[X,X]_T^{(k)}}\sqrt{[Y,Y]_T^{(k)}}}. \tag{6}$$

While there are many proposed methods (e.g. pre-averaging, realized kernels, etc.) for correcting the bias of the *RC* and *RCorr* estimators, we retain here only the subsample estimator as an indicative bias-reduced estimator. Our purpose is not to obtain bias correction but to obtain identification bounds on the *RC* and *RCorr*. Proper identification bounds would ideally, but not necessarily, include the *RC* and *RCorr* estimators as well as the the subsampled and other bias-corrected *RC* and *RCorr* estimators.

3. Identification Bounds of Realized Covariance/Correlations

Partial identification analysis departs from traditional robust statistics, which also deals with inference in the presence of data errors or problems. Robust estimation considers the stability and sensitivity of the estimators when the underlying data distribution deviates from the distribution used in the assumed model, and the objective of robust estimation is to guard against worst outcomes *ex ante*. Partial identification analysis considers these data problems in an *ex-post* setting by giving the range of the possible values of the parameter of interest given what is known about the empirical distribution. The narrower the bounds, the more information they provide.

We deem identification analysis to be well-suited for application to *RC* and *RCorr* given that they are *ex-post* or *realized* measures. However, the bounds derived in Horowitz and Manski (2006) and Horowitz and Manski (1995) are made for the case of a static framework, where time-series effects are not considered. Time dependencies would complicate the analysis of the data problems since the errors in observations would also affect the next period's observations (see for example Chen and Liu, 1993, and Tsay, Pena, and Pankratz, 2000). Fortunately for our case, *RC* and *RCorr* are derived under the framework of assuming that the returns follow Itô stochastic processes, with i.i.d. increments (Brownian motion increments are assumed). It then becomes reasonable to use identification analysis in a such a framework.

To investigate the effect of different data problems on the bounds, we derive the bounds under three cases: (i) asynchronous data without microstructure noise, (ii) synchronous data with microstructure noise, and (iii) asynchronous data with microstructure noise. While only the last case provides realistic bounds to *RC*, the first two cases allow us to observe the marginal effect of each data problem alone.

3.1 Bounds Due to Asynchronicity

Let the log returns be given as $r^X_{\tau_i} = (X_{\tau_i} - X_{\tau_{i-1}})$ and $r^Y_{\tau_i} = (Y_{\tau_i} - Y_{\tau_{i-1}})$, where τ_i, $i = 1,\ldots,M_N$ are points on a regular time grid within a day. When there are no observations within the vicinity of τ_i, we consider the data to be "missing" at τ_i. We term this vicinity as "tolerance" and it is a tuning parameter that determines the width of the bounds (see Section 3.3 for further discussion). We define four states of "missingness" and let integer valued random variable Z_i represent the state of "missingness" of the data. $Z_i = 1$ implies all components (both $r^X_{\tau_i}$ and $r^Y_{\tau_i}$) are observed. $Z_i = 2$ if $r^X_{\tau_i}$ is missing (i.e. X_{τ_i} or $X_{\tau_{i-1}}$ or both are

missing) while $r^Y_{\tau_i}$ is observed. $Z_i = 3$ if $r^Y_{\tau_i}$ is missing (i.e. Y_{τ_i} or $Y_{\tau_{i-1}}$ or both are missing) but $r^X_{\tau_i}$ is observed. $Z_i = 4$ is when both $r^X_{\tau_i}$ and $r^Y_{\tau_i}$ are missing.

Let the price pair be given by random vector $V_i = (r^X_{\tau_i}, r^Y_{\tau_i})$. Assume V_i is a discrete random variable with support $\{v_j : j = 1,\ldots,J\}$. Define $\pi_z = P(Z_i = z)$ and $p_{zj} = P(V_i = v_j \,|\, Z_i = z)$. The sampling process identifies π_z and p_{1j} for all $j = 1,\ldots,J$. For $z = 2, 3, 4$, p_{zj} is unidentified due to missingness of data, but it obeys certain restrictions. First of all, $p_{zj} \geq 0$, $z = 2,3,4$ and $\Sigma_{j=1}^{J} p_{zj} = 1$, $z = 2,3,4$. For $z = 2$, r^Y is observed. Let the support of r_Y be given by Λ_2, and suppose there are K_2 distinct points in Λ_2. Let q_{2k_2} denote the probability of point $k_2 \in \Lambda_2$ conditional on $Z = 2$. Then $\Sigma_{j:\, v_j \in k_2} p_{2j} = q_{2k_2}$, $k_2 = 1,\ldots,K_2$, where q_{2k_2} is the marginal probability of $r^Y_{\tau_i}$ in state 2.

The same applies to $z = 3$, where this time r^X is observed and r^Y is missing. Let Λ_3 refer to the support of r^X. q_{3k_3} denotes the probability of point $k_3 \in \Lambda_3$ conditional on $Z = 3$. Then $\Sigma_{j:\, v_j \in k_3} p_{3j} = q_{3k_3}$, $k_3 = 1,\ldots,K_3$, where q_{3k_3} is the marginal probability of $r^X_{\tau_i}$ in state 3.[2]

The probability mass function of the returns vector is then given by

$$P(V_i = v_j) = \sum_{z=1}^{4} \pi_z p_{zj}.$$

The upper and lower bounds can be obtained by maximising and minimising the realized covariance or correlation with respect to the unknown probabilities (p_{2j}, p_{3j} and p_{4j}) and computed using V_i over its support v_j, $j = 1\ldots,J$ subjected to the above given constraints.

3.1.1 Estimation of Bounds

The relevant optimisation problem for estimating bounds of realized covariance is given as[3]

$$\text{maximize (minimize)}_{p_{zj}:z\geq 2;j=1,\ldots,J} : \sum_{\tau_i, i\in[0,M_N]} \sum_{z=1}^{4} \sum_{j=1}^{J} \pi_z p_{zj} r^X_{\tau_i} r^Y_{\tau_i} \tag{7}$$

subject to:

2 In reality, when π_2 and π_3 are small, consistent estimation of q_{2k_2} and q_{3k_3} is difficult. However as argued by HOROWITZ and MANSKI (2006), the effects of such imprecision are limited by the multiplication with the small π_2 and π_3.

3 Here maximising the objective function gives the upper bound and minimising the objective function gives the lower bound. This is not to be confused with a max-min optimisation problem.

$$p_{zj} \geq 0,\ j = 1,\ldots,J,$$

$$\sum_{j\in[0,J]} p_{zj} = 1,\ z = 2,3,4,$$

$$\sum_{j:v_j\in k_z} p_{zj} = q_{zk_z},\ z = 2,3;\ k_z = 1,\ldots,K_z.$$

For realized correlation, the optimisation problem is:

$$\text{maximize (minimize)}_{p_{zj}:\, z\geq 2;\, j=1,\ldots,J}:$$

$$\frac{\sum_{\tau_i, i\in[0,M_N]} \sum_{z=1}^{4}\sum_{j=1}^{J} \pi_z p_{zj} r_{\tau_i}^X r_{\tau_i}^Y}{\sqrt{\sum_{\tau_i, i\in[0,M_N]} \sum_{z=1}^{4}\sum_{j=1}^{J} \hat{\pi}_z p_{zj} r_{\tau_i}^X r_{\tau_i}^X} \times \sqrt{\sum_{\tau_i, i\in[0,M_N]} \sum_{z=1}^{4}\sum_{j=1}^{J} \hat{\pi}_z p_{zj} r_{\tau_i}^Y r_{\tau_i}^Y}} \tag{8}$$

subject to:

$$p_{zj} \geq 0,\ j = 1,\ldots,J,$$

$$\sum_{j\in[1,J]} p_{zj} = 1,\ z = 2,3,4,$$

$$\sum_{j:v_j\in k_z} p_{zj} = q_{zk_z},\ z = 2,3;\ k_z = 1,\ldots,K_z.$$

The data sample allows identification of π_z (for $z = 1,\ldots,4$), p_{1j}, q_{2k_2} and q_{3k_3}. The empirical estimators are

$$\hat{\pi}_z = \frac{1}{M_N}\sum_{i=1}^{M_N} I(Z_i = z),\ z = 1,\ldots,4, \tag{9}$$

$$\hat{p}_{1j} = \frac{1}{\hat{\pi}_1 M_N}\sum_{i=1}^{M_N} I(V_i = v_j, Z_i = 1),\ j = 1,\ldots,J, \tag{10}$$

and for $z = 2,3$,

$$\hat{q}_{zk_z} = \frac{1}{\hat{\pi}_z M_N}\sum_{i=1}^{M_N} I(V_i \in k_z, Z_i = z),\ z = 2,3. \tag{11}$$

When $\hat{\pi}_1 = 0$, the right hand side of (10) is defined to be zero, and when $\hat{\pi}_z = 0$ for $z = 2, 3$, the right hand side of (11) is defined to be zero. The upper and lower bounds of *RC* and *RCorr* are then obtained by replacing π_z, p_{1j}, q_{2k_2} and q_{3k_3} with $\hat{\pi}_z, \hat{p}_{1j}, \hat{q}_{2k_2}$ and $\hat{q}_{3k_3}$ and solving (7) and (8) respectively.

Analytical solutions of (7) and (8) are not possible[4] hence we use numerical techniques to obtain the bounds. To achieve global optimization, we use the genetic algorithm[5] in MATLAB to solve the nonlinear programming problem, and then refine the optimisation locally using constraint optimisation. Optimization at each time point (per day) is approximately 10 minutes for realized covariance using a 2.66 GHz Intel computer, and 15 minutes per time point for realized correlation. Asymptotically valid confidence intervals for the bounds may be obtained using bootstrap (see HOROWITZ and MANSKI, 2006, for details), but due to the huge computational time required, we do not do this here.

3.2 Bounds Due to Microstructure Noise

To examine the effects of microstructure noise alone, let us assume that the degree of asynchronicity is negligible. Let the product of the price pair be given by random variable $W_i = r^X_{\tau_i} \times r^Y_{\tau_i}$. Let ε_{W_i} be a random variable representing microstructure noise that has an unknown distribution. The inferential problem is then that the true latent martingale price returns W_i^* is not observed but W_i which is contaminated by ε_{W_i} with probability p. Let F_W indicate the cdf of W_i. Then the sampling process only allows for the identification of F_W

$$F_W = (1-p)F_{W^*} + pF_{\varepsilon_W}, \tag{12}$$

where F_{W^*} is the distribution of the observations that belong in the efficient price distribution and F_{ε_W} is the distribution of noise.[6]

If the occurrence of microstructure noise ε_W can be assumed to be random and independent of the sampling process, then inference on F_W can be assumed to be equivalent to inference on F_{W^*}. This is termed "contaminated sampling".

4 HOROWITZ and MANSKI (2000) derived the analytical solution for the optimization problem in the case of binary outcomes. Our problem here is however more complex.

5 HOROWITZ and MANSKI (2006) found that the genetic algorithm performs the optimization faster, but alternative global optimisation techniques such as simulated annealing may be used.

6 While the additive error/noise model is commonly used in realized measures literature for deriving biases in the continuous framework, we use this discrete error model here to help us conduct our analysis. This is only possible since we are working solely in the discrete time framework.

When this independence assumption is not made, it is termed the "corrupted sampling" model and is the more general model (i.e. wider bounds under a corrupted sampling model than under contaminated sampling, see Horowitz and Manski, 1995). Both of these error models are commonly assumed in robust statistics. For example, the contaminated sampling scheme is assumed in obtaining the influence function in robust statistics, while the corrupted sampling scheme is assumed in high-breakdown estimation (see a review by Pavel Čížek and Härdle, 2006). We consider here both the contaminated and corrupted sampling schemes.

Horowitz and Manski (1995) showed that for parameters which respect stochastic dominance, sharp bounds can be obtained under corrupted and contamination sampling. *RC* respects stochastic dominance with respect to W_i, for $W_i \in R^*$ where R^* is the extended real line. This means that $RC(F^1_W) \geq RC(F^2_W)$ when $F^1_W \leq F^2_W$.

3.2.1 Estimation of Upper Bound of p

While we need not assume a value for p, the analysis requires that an upper bound of p can be estimated or known a priori. Denote this upper bound as $\lambda < 1$ such that $\lambda \geq \max(\lambda^X, \lambda^Y)$, for $p_X \leq \lambda^X < 1$ and $p_Y \leq \lambda^Y < 1$, where p_X and p_Y are contamination probabilities of X and Y respectively, and λ^X and λ^Y are their corresponding upper bounds. To estimate the level of microstructure noise, we use Eq (14) of Bandi and Russell (2008) who derive the bias in realized variances (RV) due to microstructure noise[7]:

$$E(\widehat{RV} - RV) = M_N E(\varepsilon_X^2) \tag{13}$$

where $E(\varepsilon_X^2)$ can be estimated consistently using Theorem 2 (Bandi and Russell, 2008):

$$\frac{1}{M_N}\sum_{j=1}^{M_N} r_j^{2,X} \xrightarrow{p} E(\varepsilon_X^2) \text{ as } M_N \to \infty. \tag{14}$$

This result is also obtained in Zhang, Mykland, and Aït-Sahalia (2005) and the intuition is that at very high sampling frequencies, the realized variance is a

7 While this bias term is obtained under the assumption that microstructure noise follows a white noise process that is independent of the price process, we can use this bias term to obtain a rough estimate of the percentage of microstructure noise in the data.

consistent estimator of the variance of microstructure noise. BANDI and RUSSELL (2008) propose using an average of bias estimation over n samples (rolling windows can be used to allow time-variation in bias), but we will use the maximum estimated bias over the period to estimate the upper bound of microstructure noise λ^X and λ^Y separately and take λ such that $\lambda \geq \max(\lambda^X, \lambda^Y)$.

3.2.2 Estimation of Bounds

Under Proposition 4 in HOROWITZ and MANSKI (1995), the bounds of RC under contamination, $RC(W_{cont})$, is given by

$$RC(W_{cont}) \in \left[\sum_{i=1}^{M_N}\{W_{i,L}\}, \sum_{i=1}^{M_N}\{W_{i,U}\}\right] \tag{15}$$

where $W_{i,L}$ and $W_{i,U}$ are random variables drawn from distributions

$$L[-\infty,t] \equiv \begin{cases} \dfrac{F_W[-\infty,t]}{1-\lambda}, & \text{if } t < F_W^{-1}(1-\lambda) \\ 1, & \text{if } t \geq F_W^{-1}(1-\lambda) \end{cases} \tag{16}$$

and

$$U[-\infty,t] \equiv \begin{cases} 0, & \text{if } t < F_W^{-1}(\lambda) \\ \dfrac{F_W[-\infty,t]-\lambda}{1-\lambda}, & \text{if } t < F_W^{-1}(\lambda) \end{cases} \tag{17}$$

respectively.

For corrupted sampling, let $\delta_{-\infty}$ and δ_{∞} be the limiting probability measures on $\mathbb{W}$ at $-\infty$ and ∞ respectively, then the bounds are

$$RC(W_{corr}) \in \left[\sum_{i=1}^{M_N}(W_{i,L_2}), \sum_{i=1}^{M_N}(W_{i,U_2})\right] \tag{18}$$

where $L_2 = (1-\lambda)L + \lambda\delta_{-\infty}$ and $U_2 = (1-\lambda)U + \lambda\delta_{\infty}$.

Consistent estimation of the bounds under contaminated sampling can be obtained by $[0,1] \cap [(\hat{F}_W - \hat{\lambda})/(1-\hat{\lambda}),\ \hat{F}_W/(1-\hat{\lambda})]$ and the bounds under corrupted sampling by $[0,1] \cap [\hat{F}_W - \hat{\lambda},\ \hat{F}_W + \hat{\lambda}]$, where $\hat{F}_W$ is estimated using the

empirical cdf of W. The corresponding bounds for *RCorr* under the corrupted ($RCorr(W_{corr})$) and contaminated ($RCorr(W_{cont})$) schemes are then obtained by dividing the bounds obtained in *RC* by the realized volatilities of X and Y.

3.3 Overall Bounds and Estimation Issues

We assume that the effects of asynchronicity and microstructure noise are independent, separate data problems, and obtain overall bounds on *RC* and *RCorr* by summing the individual bounds. Since the bounds obtained due to microstructure noise alone are symmetric, we add half of the bound due to noise to the upper bound due to asynchronicity, and subtract half of the bound due to noise from the lower bound due to asynchronicity to obtain the overall bounds on *RC* and *RCorr*, i.e.

$$U_{\text{overall}} = U_{\text{asyncronicity}} + \frac{1}{2(U_{\text{noise}} - L_{\text{noise}})}$$

and

$$L_{\text{overall}} = L_{\text{asyncronicity}} - \frac{1}{2(U_{\text{noise}} - L_{\text{noise}})},$$

where U_p and L_p refer to upper and lower bounds due to p effect.

In reality, these two effects may not be independent and may be instantaneously causal for each other. Then the actual overall bounds may be narrower than what we estimate here, given that we are measuring these effects ex-post – some of the effects of asynchronicity would already be included when estimating the bounds due to microstructure noise and vice versa. By adding the bounds of the two effects, we would likely obtain a more conservative estimate of the overall bounds. From our empirical application (see Section 5), we find that this method gives rather tight bounds, hence we do not delve further into the issue of instantaneous causality of the two effects.

Estimation of the bounds encounters several issues: First, for computational tractability in estimating bounds due to asynchronicity, the support of the returns is assumed to be finite, but in theory the support is $[-\infty, \infty]$. What this entails is that if there is a large proportion of missingness that in reality would lie outside the support of the observed returns, the width of the bounds would be underestimated.

Second, the estimation of bounds is subjected to error due to discretization of the support. In our applications, we used 10-by-10 bins for approximating the

two-dimensional support to obtain bounds due to asynchronicity. Finer discretization would reduce the degree of error but would require greater computational power.

Third, the overall width of the bounds is subject to two tuning parameters. For bounds due to asynchronicity, the "tolerance" or vicinity of the grid for assigning "missingness" (i.e. if no observations are observed within x seconds before the sampling grid, the observations are considered missing) is the tuning parameter. The smaller the tolerance chosen, the wider the bounds. At high tolerance levels and low asynchronicity, the bounds narrow towards the previous-tick *RC*. We consider a base case of 15 seconds for the 5 minutes *RC* and *RCorr*, or 5% of the sampling interval. For bounds due to microstructure noise, the tuning parameter is the upper bound of microstructure noise λ. The higher the estimated or assumed λ, the wider the bounds. Simulations in Section 4 will check the efficacy of these bounds and their sensitivity to the tuning parameters.

3.4 Forecasting Correlation and its Bounds Using HAR

Besides providing an identification region for an estimator, bounds could be useful in providing prediction of the region where the parameter could potentially lie in. The forecasting of *RC* is a challenging issue because it is usually made in the context of forecasting the entire variance-covariance matrix, and in this multivariate dimension, the forecast model engages in problems with parameter proliferation and ensuring the positive definiteness of the variance-covariance matrix. Hence we consider forecasting realized correlations rather than realized covariance. Inherent in conducting the forecast of its bounds is the assumption that the degree of asynchronicity and microstructure noise remains constant.

We forecast realized correlations using the Heterogeneous Autoregressive (HAR) Model, a parsimonious model for modeling long-memory processes[8]. Realized correlations tend to empirically exhibit a high degree of persistence, hence HAR is a simple and suitable forecasting model (see for example CORSI and AUDRINO, 2007, and VORTELINOS, 2010). The HAR model for forecasting the 1-step ahead realized correlations is given by

$$RCorr_{t+1}^{(d)} = c + \beta^{(d)} RCorr_t^{(d)} + \beta^{(w)} RCorr_t^{(w)} + \beta^{(m)} RCorr_t^{(m)} + \varepsilon_{t+1} \qquad (19)$$

8 Refer to CORSI (2003) and CORSI (2009) for more information about HAR models.

where $RCorr_t^{(d)}$, $RCorr_t^{(w)}$ and $RCorr_t^{(m)}$ are the daily, weekly and monthly realized correlations respectively. We obtain the weekly $RCorr^{(w)}$ and monthly $RCorr_t^{(m)}$ by taking averages of the last 5 and 20 days *RCorr* respectively.

We extend the application of the HAR model to forecast the bounds of realized correlations, i.e. we replace *RCorr* by the bounds $U_{overall}$ or $L_{overall}$[9]. Estimation of (19) is made via ordinary least squares. To achieve multi-period forecasting, we use iterative forecasts to obtain the h-step ahead forecast.

4. Simulations

We simulate a bivariate Brownian process with zero drift and constant covariance using the Euler discretization scheme to obtain prices at 1 second intervals.[10] The price process is then $dP = \Sigma_t^{1/2} dB$, where $\Sigma_t^{1/2}$ is the Cholesky factorisation of the covariance matrix Σ_t and B is a (2×1) vector of independent standard Brownian motions. The integrated covariance is then given by

$$IC = \int_0^1 \Sigma_t dt.$$

To obtain non-synchronous price pairs, we simulate durations using independent Poisson processes with constant intensities ϕ_X and ϕ_Y. An additive noise process in the form of Gaussian white noise

$$u \sim N\left(0, \begin{pmatrix} \omega_{XX} & \omega_{XY} \\ \omega_{XY} & \omega_{YY} \end{pmatrix}\right)$$

is added to the price processes.

We use simulation values

$$IC = \begin{pmatrix} 1.0e^{-4} & 1.2e^{-4} \\ 1.2e^{-4} & 2.5e^{-4} \end{pmatrix}$$

which are close to what is observed empirically. Multivariate normal contemporaneously correlated noise is added with $\omega_{XX} = 1.0e^{-5}$, $\omega_{YY} = 2.5e^{-5}$ and $\rho = -0.1$. We use this simple simulation setup to evaluate the estimated bounds with respect to the *RC* estimators and the true *RC*.

9 We forecast the bounds directly, and not obtain the bounds of the forecast *RCorr*.
10 For simplicity, we do not consider leverage effects.

4.1 Bounds' Sensitivity to Asynchronicity

There does not exist at present a standard measure for the degree of asynchronicity, hence we use simple simulated trials to observe the effect of the tuning parameter on the estimated bounds due to asynchronicity. Figure 1 plots the width of 5 minutes *RC* bounds due to asynchronicity against tolerance using simulated data with asynchronicity alone for 3 cases: (i) $\phi_X = 1$, $\phi_Y = 1/5$ (durations of X and Y are 1 and 5 seconds respectively), (ii) $\phi_X = 1$, $\phi_Y = 1/10$ (durations of X and Y are 1 and 10 seconds respectively) and (iii) $\phi_X = 1/5$, $\phi_Y = 1/10$ (durations of X and Y are 5 and 10 seconds respectively). Cases (i) and (iii) have the same approximate degree of asynchronicity (differences in durations are 4 and 5 seconds respectively), while case (ii) has the greatest degree of asynchronicity (with difference in durations of 9 seconds).

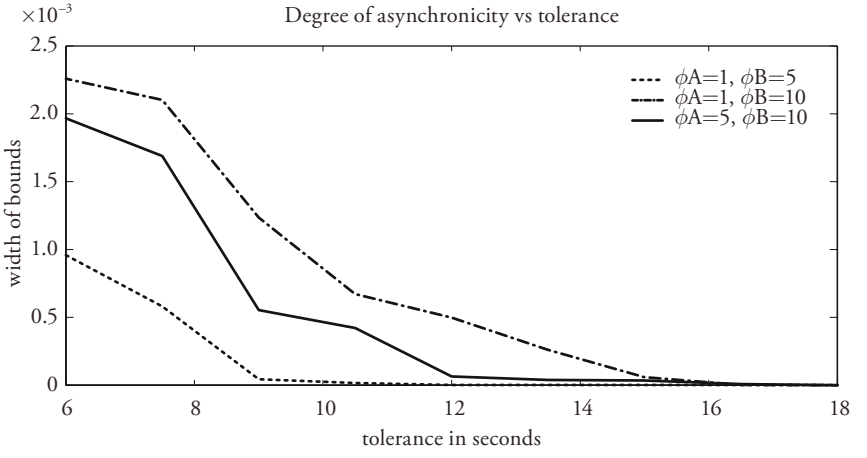
Figure 1 shows that as expected, for all three cases, the width of bounds decreases as tolerance increases (i.e. as the vicinity of what we consider 'missing' becomes narrower, the estimated bounds become wider). Second, bounds of case (ii) are wider than bounds of cases (i) and (iii), and width of case (iii) is larger than case (i), which confirms that the greater the degree of asynchronicity, the larger is the width of the bounds. Third, although case (i) and case (iii) have approximately the same degree of asynchronicity, the width of case (iii) is much larger than case (i), which suggests that the lower the intensities of the arrival rates (i.e. larger durations), the larger the width of the bounds. This is intuitive since the lower the intensity, the greater the probability that no observations will be observed within the vicinity or "tolerance" of the grid.

4.2 Bounds' Sensitivity to Level of Microstructure Noise

We now turn to the sensitivity of bounds due to the effect of microstructure noise alone. Figure 2 shows the 5 minutes *RC* and its bounds with synchronous observations for 100 simulated days using constant noise-signal ratios of 10% ($p = 0.1$, Fig. 2a) and 20% ($p = 0.2$, Fig. 2b). The horizontal solid line close to the x-axis is the true *RC*, the solid fluctuating line is the previous-tick *RC* and the dotted line that exhibits lesser fluctuations than previous-tick *RC* is the *ssRC*. The *ssRC* tends to lie consistently under the true *RC*, hence the *ssRC* still incurs a bias but it is much less biased than the previous-tick *RC*.

The inner and outer bounds are the bounds under contaminated and corrupted sampling respectively (bounds under the corrupted sampling scheme are wider than those under the contaminated sampling scheme, see Section 3.2). The bounds provide total coverage of the previous-tick *RC* and *ssRC*. For the

Figure 1: Plots of the Width of 5 min *RC* Bounds due to Asynchronicity against Tolerance (Vicinity of Grid) Using Simulated Data without Noise



Notes: Three cases are plotted here: (i) $\phi_X = 1$, $\phi_Y = 1/5$ (1 and 5 sec durations) (dotted line), (ii) $\phi_X = 1$, $\phi_Y = 1/10$ (1 and 10 sec durations) (dash-dot line) and (iii) $\phi_X = 1/5$, $\phi_Y = 1/10$ (5 and 10 sec durations) (solid line).

true *RC* (solid constant line), the bounds under the corrupted sampling scheme provide 100% coverage, but under contaminated sampling, it is less that 100%. This implies that the bounds under contaminated sampling are not as reliable in the inclusion of the true *RC*. This also applies when adding a 20% microstructure noise (Figure 2b).

Finally, as expected, the bounds widen as the level of additive microstructure noise increases. This is observed in Figure 2 where the bounds in (b) are wider than those in (a).

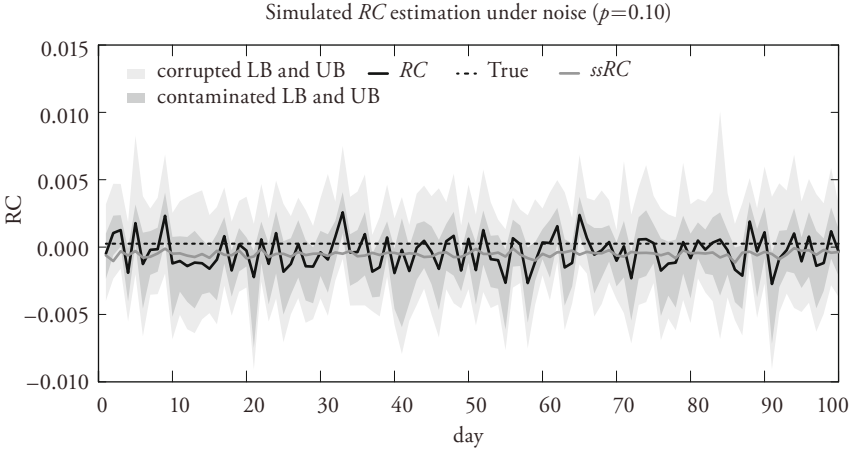
4.3 Overall Bounds and Sampling Frequency

Figure 3 shows simulations under 10% microstructure noise ($p = 0.1$) and asynchronicity of $\phi_X = 1/5$ and $\phi_Y = 1/10$ (X and Y have 1 and 10 minute durations respectively). The estimations of the 5 minutes *RC* is given in Figure 3a while the 1 minute *RC* is given in Figure 3b. The tolerance is set at 15 seconds (i.e. if no observations are within 15 seconds of the grid, the observation is considered "missing").

Figure 2: True *RC* (Dotted Line), Previous Tick *RC* (Solid Black Line), *s* *sRC* (Solid Grey Line) and Estimated Bounds under Microstructure Noise alone at 5 Minutes Sampling Intervals

Additive microstructure noise is at (a) 10% and (b) 20% noise-signal levels. Estimated bounds assume the corrupted sampling scheme (light grey area) and contaminated sampling scheme (darker grey area).

(a) 10% noise-signal ratio ($p = 0.1$)



(b) 20% noise-signal ratio ($p = 0.2$)

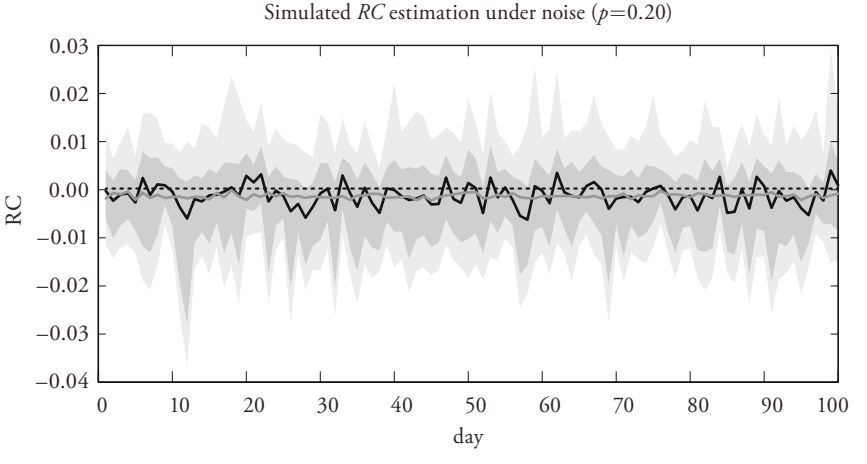
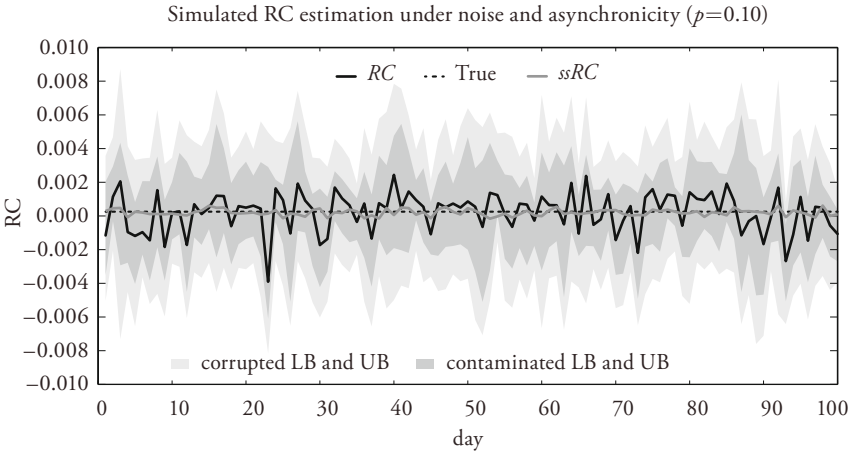
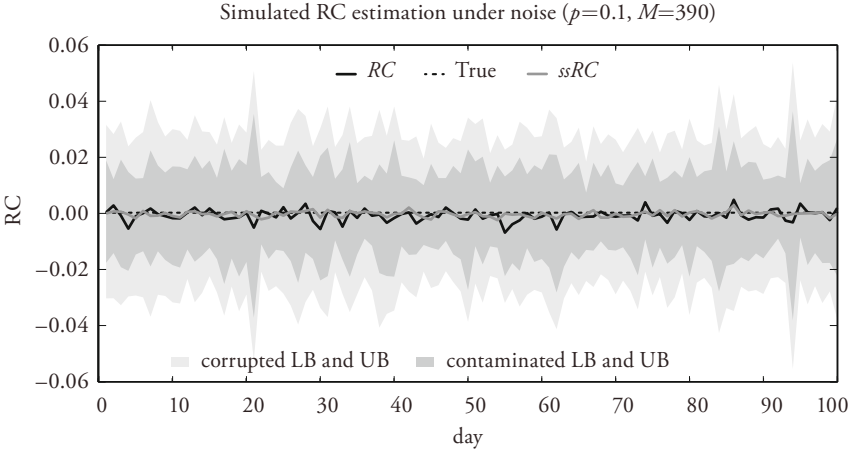


Figure 3: Estimated *RC* and overall bounds for 100 simulated days using additive microstructure noise of 10% ($p=0.1$), $\phi_X=1/5$, $\phi_Y=1/10$ at (a) 5 minutes sampling intervals and (b) 1 minute sampling intervals

(a) 5 minute *RC* and corresponding bounds



(b) 1 minute *RC* and corresponding bounds



For both cases, the bounds provide complete coverage under both types of sampling (except for one day in Figure 3a under contaminated sampling). The bounds under 1 minute sampling (Figure 3b) are wider than those under 5 minutes sampling (Figure 3a). This is expected as under 1 minute sampling, there are more points on the sampling grid and hence a greater probability for "missingness" to occur.

5. Data

For our empirical study, we use the NYSE TAQ quote data of Citigroup (c) and JP Morgan Chase (jpm) for the year 2007 (2 Jan 2007 to 31 Dec 2007), a total of 251 days. The mid-quotes during regular trading hours between 9.30 and 16.00 are extracted. For multiple quotes within the same second, the average mid-quote is used.

Figure 4: Previous-Tick Realized Covariance Signature Plot

The realized covariance tends to zero as sampling frequency increases due to the Epps effect (observed here between 1–5 minutes). For previous-tick realized covariance, the effect of microstructure noise at the highest sampling frequency is also observed.

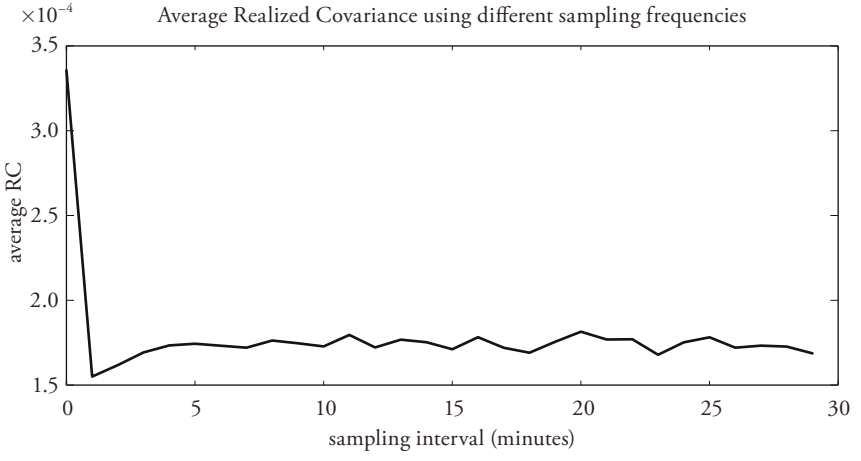


Figure 4 shows the signature plot of realized covariance, which is a plot of the average realized covariance using different sampling frequencies. The x-axis gives the sampling frequency in minutes (i.e. Δt, the interval between each sampling).

The smallest sampling interval used is one second. The y-axis gives the average estimated previous-tick *RC* over the sample period. As is characteristic of realized volatility signature plots, there is a large upward slope at very high sampling frequencies due to the effect of microstructure noise[11]. This is in line with the asymptotic theory for microstructure noise by Bandi and Russell (2008). However for realized covariance, an additional Epps effect is observed between 2–5 minutes sampling interval, where there is a bias towards zero as the sampling frequency increases. The *RC* stabilises at the 5 minutes sampling interval, hence we will use the 5 minutes *RC* and *RCorr*.

Figure 5 shows the realized covariance and correlation of c and jpm at 5 minutes sampling frequency using the previous tick *RC* and subsampled estimator *ssRC*. The graphs show sharp increases in both covariance and correlations in the second half of 2007 due to the effects of the credit crisis. The previous-tick *RC* and *ssRC* also produce substantially different values for covariance and correlations. This difference is more noticeable for realized correlations due to scaling effects.

5.1 Results

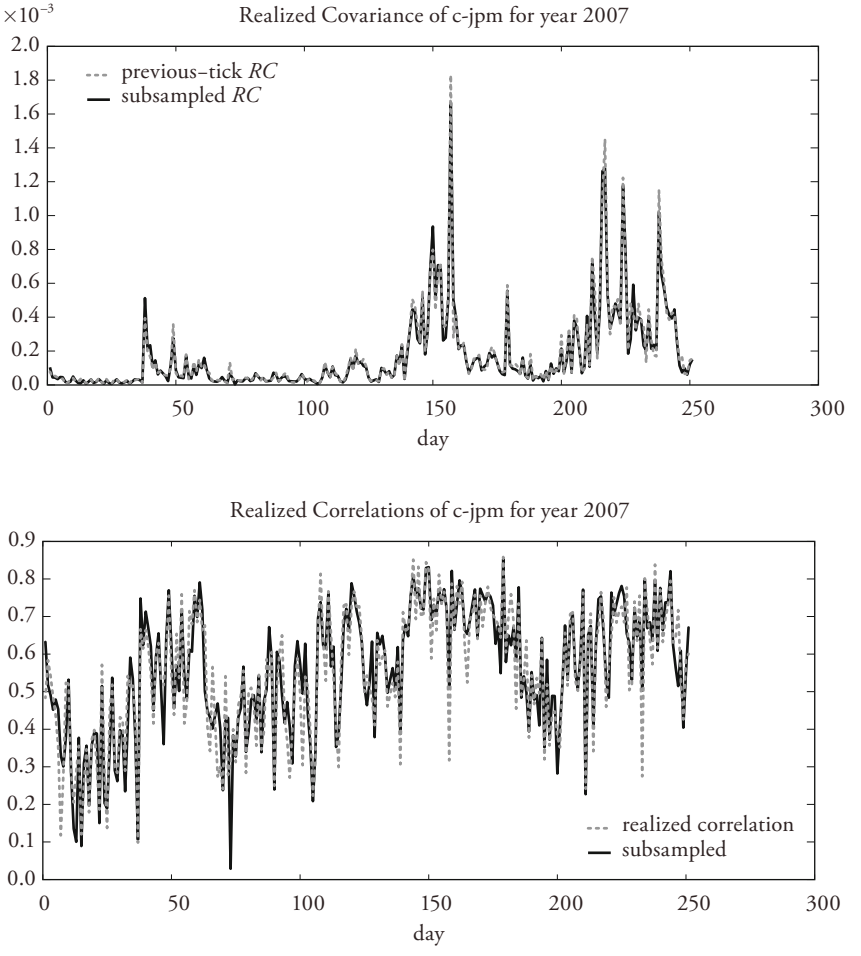
5.1.1 Estimated Bounds Due to Asynchronicity

We first present the results for bounds due to asynchronicity of data alone. If there are no observations within a "tolerance" vicinity of the sampling grid, we consider the observation as latent or "missing". For our case, we set the tolerance level to 5% of the sampling interval Δt. Since we use 5 minutes sampling frequency, the state of "missingness" is conferred if there is no observed data within the last 15 secs of the sampling grid.

Figure 6 shows the average estimated probabilities obtained that are used to estimate the maximum bounds of realized covariance. The x- and y-axis give the jpm and c returns respectively, while the z-axis gives the average estimated probabilities over the sample period. Figure 6a graphs the average observable probabilities p_{1j} (i.e. $z=1$). It has a cone shape that peaks at the centre where returns are approximately 0 for both c and jpm. This is expected as 5 minutes intraday returns are relatively small with approximately zero mean. Figure 6b shows the average estimated probabilities at $z=2$, where returns of c are observed but returns of jpm are missing. Peaks tend to be along the x-y axis, which gives the

11 At very high sampling frequencies, only microstructure noise is measured, see for example Figure 1 in Bandi and Russell (2008).

Figure 5: Plots of the Previous-Tick (Dotted Line) and the Subsampled (Solid Line) Realized Covariance (Top) and Realized Correlations (Bottom) of The Sample Period Using 5 Minutes Sampling Frequency



maximum covariance. Figure 6c gives the estimated probabilities at $z=3$, where returns for jpm are observed but returns of c are missing. Since c is more frequently traded than jpm, π_3 tends to be small. Under $\pi_3=0$, the probabilities do not enter the optimisation problem and are relegated to the first bin (smallest c and jpm returns), hence giving the sharp peak at that corner. The rest of

the probability surface is rather flat with some increase at the largest value for c and positive values of jpm. Figure 6d shows the average estimated probabilities when both c and jpm returns are not observed. Again there is a peak in the first bin (smallest c and jpm returns) for days when $\pi_4 = 0$ and another peak at the opposite end where both jpm and c returns are largest, which gives the largest covariance.

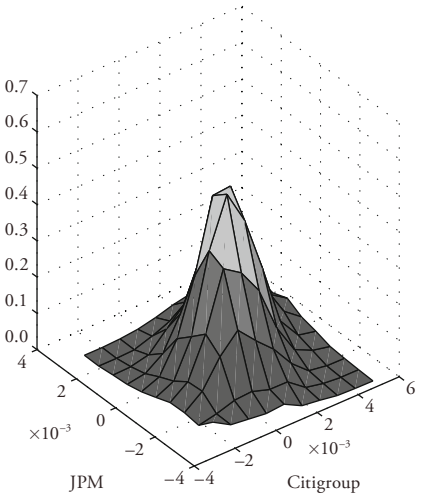
Figure 7 shows the average estimated probabilities used to estimate the minimum bounds of realized covariance. Figure 7a gives the average observable probabilities p_{1j}, which is the same as in Figure 6a. Figure 7b gives the probabilities for $z = 2$ when returns for c are observable and returns for jpm are missing. The peak occurs at the maximum value for c and the minimum value for jpm, which gives the smallest covariance. A smaller peak is observed at the largest value for jpm and smallest value for c which also decreases the value of realized covariance. The peak at the first bin (small c and small jpm) is again caused by days when $\pi_2 = 0$. Figure 7c gives the probabilities for $z = 3$. The large peak in the first bin is caused by days when $\pi_3 = 0$, and there is some small increments along the negative c values and positive jpm values which gives smaller realized covariances. Figure 7d gives the probabilities for $z = 4$ where both c and jpm returns are unobserved. Besides the peak at the first bin for days when $\pi_4 = 0$, there are two peaks opposite each other at maximum jpm returns and minimum c and vice versa, which both give the smallest realized covariance. For realized correlations, the shapes of the probability surfaces are more complicated due to the division by realized volatilities (Eq. 8), but the analysis remains similar[12].

The maximum and minimum bounds of the realized covariance and realized correlations due to asynchronous trading alone are plotted in Figure 8. As c and jpm are both highly traded stocks, the effect of asynchronicity is often small (maximum and minimum bounds are close to each other). For realized covariance, the bounds tend to widen at points of sharp increases or decreases of *RC*. This implies that for sharp increase or decrease in return of one stock, trading of the other stock tends to temporarily slow down, hence resulting in unobservability of its returns. This illustrates that asynchronicity is linked to an increase in investors' uncertainty. Similarly for realized correlations, the bounds are observed to widen around the time of the onset of the credit crisis, implying greater uncertainty during that period.

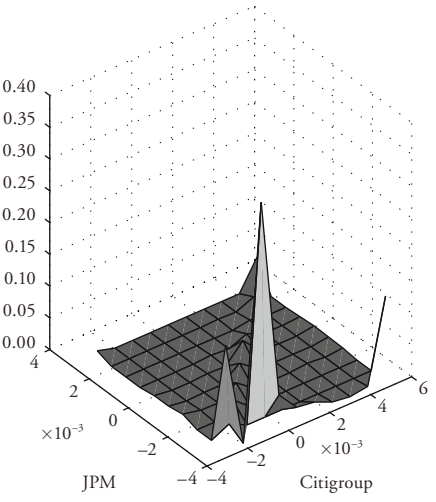
12 Graphs are available upon request from the authors.

Figure 6: Average Estimated Probabilities $\hat{p}_{zj}$ for Estimating Maximum Bounds Due to Asynchronicity for Realized Covariance

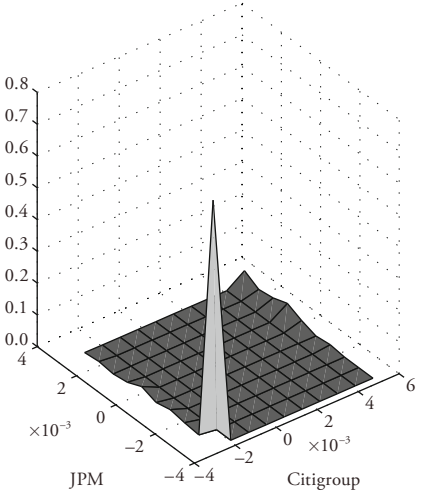
(a) $\hat{p}_{1j}$ (both jpm and c returns are observed)



(b) $\hat{p}_{2j}$ (c returns observed, jpm returns missing)



(c) $\hat{p}_{3j}$ (jpm returns observed, c missing)



(d) $\hat{p}_{4j}$ (both c and jpm returns missing)

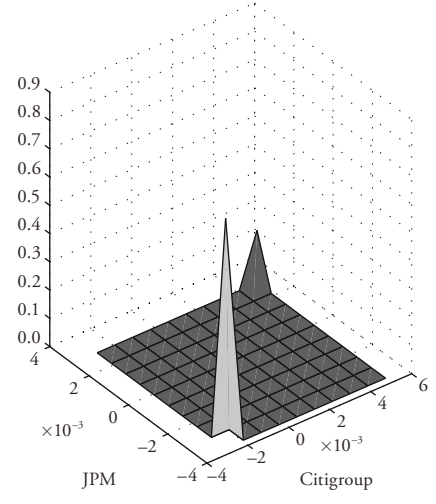
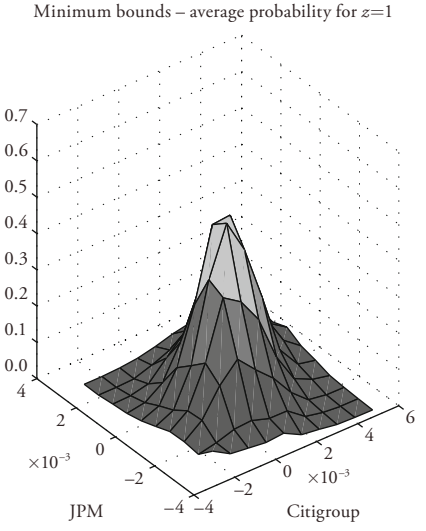
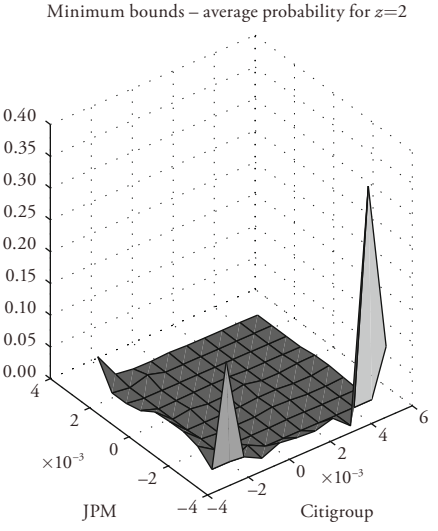


Figure 7: Average Estimated Probabilities $\hat{p}_{zj}$ for Estimating Minimum Bounds Due to Asynchronicity for Realized Covariance

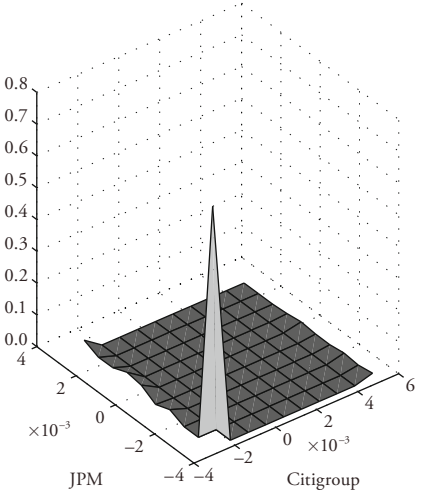
(a) $\hat{p}_{1j}$ (both jpm and c returns are observed)



(b) $\hat{p}_{2j}$ (c returns observed, jpm returns missing)



(c) $\hat{p}_{3j}$ (jpm returns observed, c missing)



(d) $\hat{p}_{4j}$ (both c and jpm returns missing)

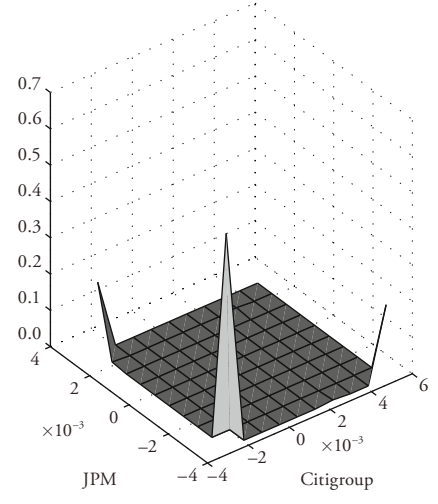
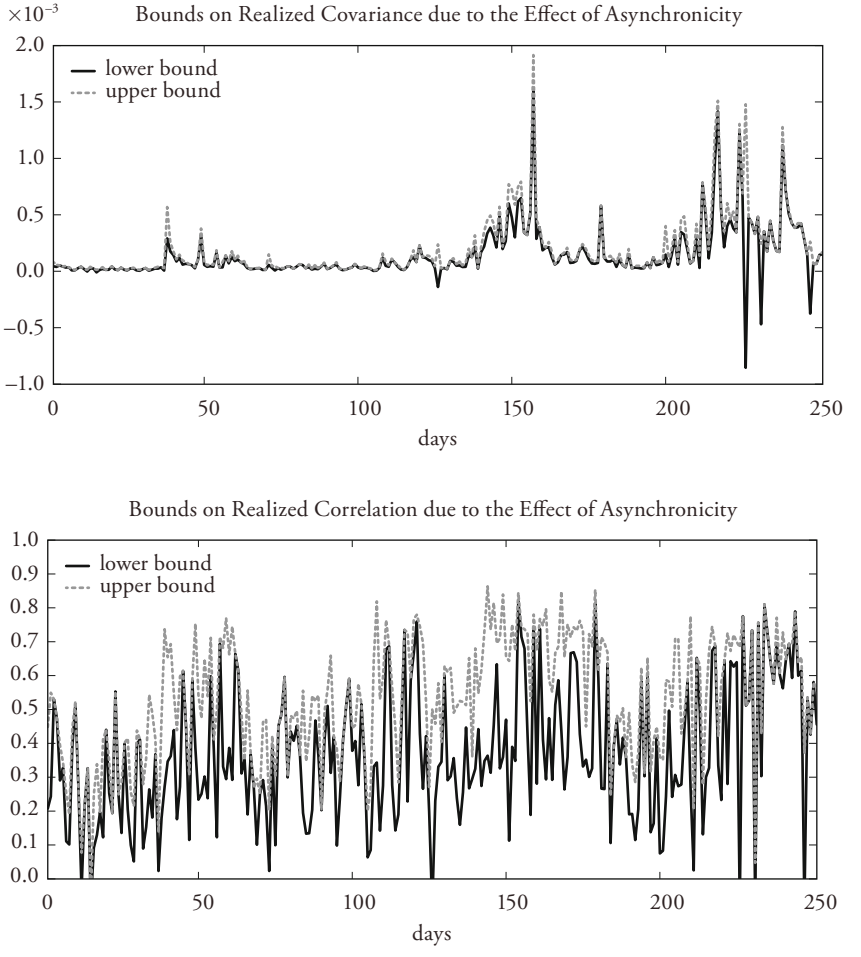


Figure 8: Bounds on Realized Covariance *RC* (top) and Correlations *RCorr* (bottom) Due to the Effect of Asynchronicity

Upper bounds are represented by grey dashed lines while lower bounds are represented by solid black lines



5.1.2 Estimated Bounds Due to Microstructure Noise

We estimate the percentage of microstructure noise as in Eq (13) and (14), and find the noise levels to be relatively low at about 1–2%, and also higher in the second half of the year. A sharp spike to over 11.5% is observed for c on 5 Nov, when Citigroup's rating was downgraded by Moodys. We interpret this spike to be the degree of deviation from the true latent efficient process and take the upper bound of noise levels λ to be 12%, in recognition that the percentage noise computed here may not be precise due to assumptions involved in obtaining the bias term. We will check the reasonableness of this upper bound later. In any case, as argued in Horowitz and Manski (1995), even if there are no obvious ways to set a firm bound, it will still be useful to analyse the sensitivity of the bounds under different error probabilities.

Figure 9 plots the width of the bounds due to effects of microstructure noise alone under corrupted sampling (solid black lines) and contaminated sampling (grey solid lines). The bounds widen dramatically in the second half of 2007 due to the onset of the credit crisis. We also include the width of the 95% confidence bands (black dotted lines) of realized covariance[13] for comparison, and find this to be rather close to that for corrupted sampling, which suggests that the confidence bands of realized covariance capture roughly captures the uncertainty due to the effects of microstructure noise.

5.1.3 Estimated Overall Bounds and Forecast

Table 1 gives the percentage coverage rate of the overall bounds on the *RC*, *ssRC*, *RCorr* and *ssRCorr* estimators for the first half of 2007 up to 30 June (day 1–124) and the second half of 2007 up to 31st December (day 125–251). The second half of 2007 will later be the out-of-sample period that is forecasted, while the first half of 2007 serves as the initial in-sample period. The coverage of the bounds under the corrupted sampling on the estimators in the first half of 2007 is 100% but the coverage of the bounds under contaminated sampling is inadequate especially for the *ssRCorr* which has a coverage rate of 82%. This implies that microstructure noise cannot be regarded as statistically independent of the efficient price process, as was also concluded by Phillips and Yu (2006) and Barndorff-Nielsen et al. (2008). Market microstructure theory also predicts the noise process to be correlated with the efficient price process (see Kalnina

13 We use the result in Barndorff-Nielsen and Shephard (2004), Equation 28, to compute the confidence bands.

Figure 9: Width of Bounds Due to Microstructure Noise for Realized Covariance

Black solid lines show width of bounds obtained under corrupted sampling, while grey solid lines show the width under contaminated sampling. The width of the 95% confidence bands (black dotted lines) are also included for comparison.

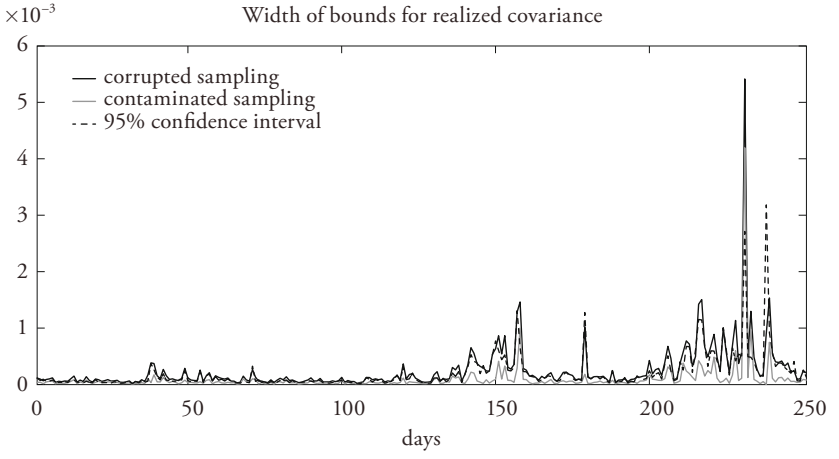


Table 1: Percentage Coverage by Overall Bounds of the Previous-Tick Realized Covariance and Correlation (*RC* and *RCorr*) and Subsampled Realized Covariance and Correlation (*ssRC* and *ssRCorr*) under Corrupted and Contaminated Sampling

	Realized Covariance				Realized Correlations			
% coverage	1st half 2007		2nd half 2007		1st half 2007		2nd half 2007	
overall bounds	*RC*	*ssRC*	*RC*	*ssRC*	*RCorr*	*ssRCorr*	*RCorr*	*ssRCorr*
corrupted	100	100	99.21	95.28	100	99.19	95.28	96.85
contaminated	100	98.39	96.06	83.46	90.32	82.26	87.40	83.46

and LINTON, 2006), hence our results here are not only unsurprising but also indicate that the bounds are tight and that the estimate of the upper bound of microstructure noise is reasonable. The coverage of the bounds during the volatile second half of 2007 crisis period is less than 100%, with some observations exceeding the upper bounds. Coverage is however still above the 95% level for the corrupted sampling scheme.

Figure 10 plots the overall bounds due to asynchronous data and microstructure noise for realized correlations under corrupted sampling. *RCorr* (black lines) and *ssRCorr* (grey lines) estimators are also graphed. It shows that the *RCorr* estimators lie closer to the upper bounds than the lower bounds. The same is observed for *RC*. This asymmetry suggests that not only is the noise process correlated with the efficient price process, its distribution in reality is also positively skewed, which causes the estimated *RC*s to lie higher than if noise were normally distributed (see for example simulation in Section 4 where this is the case)[14].

Figure 10: Overall Bounds on Realized Correlations

Grey areas show the overall bounds under corrupted sampling for microstructure noise. *RCorr* is given by the black lines and *ssRCorr* is given by the grey lines.

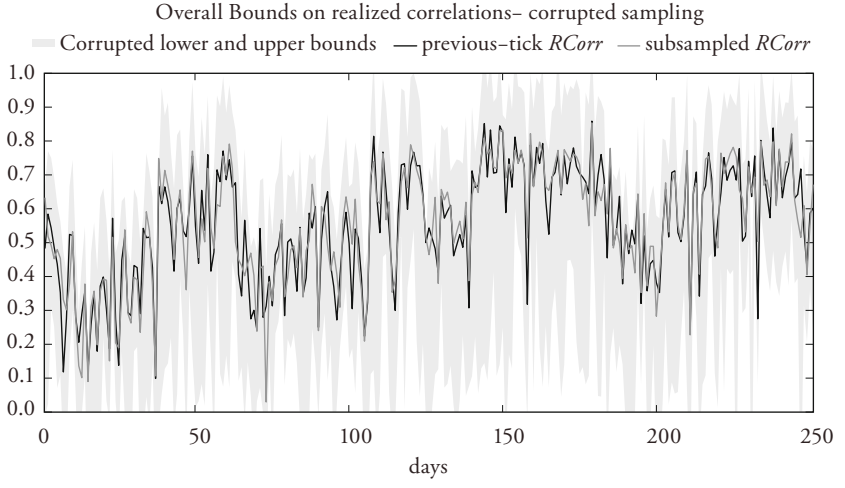


Table 2 shows the initial in-sample estimates of the HAR model for *RCorr*, *ssRCorr* and their bounds. The Newey-West standard deviations using four lags are given in brackets. The coefficients for weekly and monthly correlations are insignificant due to the small sample size used (e.g. for monthly correlations there

14 An explanation for the asymmetry in *RCorr* could lie in the fact that the measured correlation is rather high but the upper bound cannot exceed 1. This however does not explain the obvious asymmetry observed for *RC*.

are only six observations) such that the effects of long memory are harder to be captured by the model.

To obtain out-of-sample 1- and 10-step forecasts for the second half of 2007, we use rolling windows of 124 observations (approximately 6 months) to estimate the parameters. Iterated forecasts are used to obtain 10-step ahead forecasts. The predictive mean square error (PMSE) of the 1- and 10-steps forecasts are shown in Table 2. PMSE for the bounds are larger than PMSE for *RCorr* and *ssRCorr*. However for multistep forecasting, the worsening in terms of PMSE as compared to the 1-step forecast is less severe for the bounds than *RCorr* and *ssRCorr*. This feature suggests that bounds can be effectively forecasted for longer periods ahead with greater certainty as compared to point estimators. In terms of coverage, coverage at 1-step forecast is equally good as the actual coverage (see Table 1) and at 10-step forecast, although there is a slight reduction in percentage coverage, it remains above 90%.

Table 2: HAR In-Sample Estimations for Realized Correlations and Bounds and Out-Of-Sample Forecast Evaluations. Newey-West Standard Errors with 4 Lags Are Given in Parentheses.

In-sample HAR coefficient estimates								
	RCorr		*ssRCorr*		Lower Bound		Upper Bound	
c	0.1835	(0.0759)	0.1756	(0.0713)	0.0494	(0.0272)	0.2356	(0.1039)
$\beta^{(d)}$	0.2396	(0.1162)	0.3451	(0.1235)	0.1639	(0.0987)	0.3509	(0.1007)
$\beta^{(w)}$	0.4398	(0.2105)	0.2228	(0.2200)	–0.0369	(0.2469)	0.2578	(0.1869)
$\beta^{(m)}$	–0.0312	(0.2217)	0.1016	(0.2269)	0.5738	(0.4054)	0.0718	(0.2161)

PMSE (Out-of-sample)							
RCorr		*ssRCorr*		Lower Bound		Upper Bound	
1 step	10 step	1 step	10 step	1 step	10 step	1 step	10 step
0.0172	0.0263	0.0143	0.0222	0.0444	0.0487	0.0225	0.0328

Percentage coverage by forecasted bounds (Out-of-sample)					
	1 step	10 step		1 step	10 step
RCorr	96.85	93.22	*ssRCorr*	96.85	91.53

Newey-West standard errors with 4 lags are given in brackets.

5.2 Economic Significance

Identification bounds can find use in many problems in finance, and is especially interesting for risk management and portfolio allocation problems, where risk managers and portfolio managers consider these worst- and best- case scenarios given the data problems when using realized correlations. For market risk managers who are required to report the 1- and 10-day Value-at-Risk (VaR), defined as the next period's forecast loss with a certain probability, the above forecast of the identification bounds could be used instead or in addition to VaR when data problems are more severe.

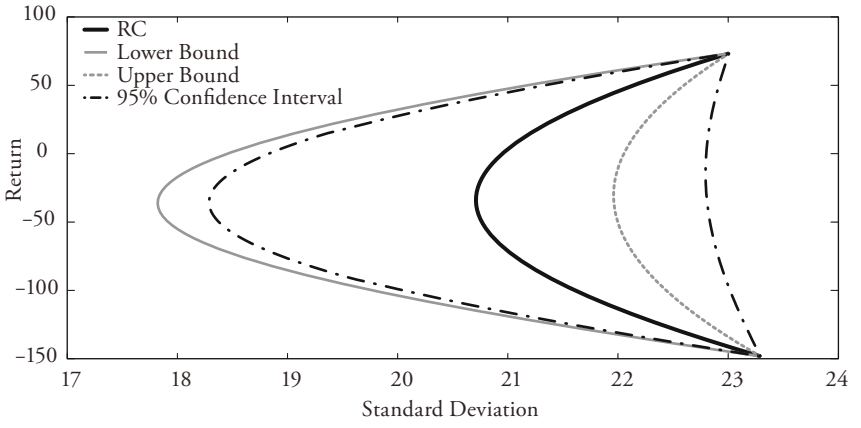
To illustrate the economic value to portfolio allocation, we consider the portfolio optimization problem of a risk-averse investor who wants to minimize the portfolio volatility targeting a certain return. His optimal portfolio is then the solution to

$$\min_{w_t} w_t' \hat{H}_t w_t \text{ s.t. } w_t' r_t = \mu_p \text{ and } w_t' \gamma = 1 \tag{20}$$

where w_t is the $n \times 1$ vector of portfolio weights for n assets at time t, $\hat{H}_t$ is the estimated conditional covariance matrix at time t, r_t is the $n \times 1$ vector of returns of the individual assets at time t, μ_p is the target return and γ is an $n \times 1$ vector of ones. By solving (20) for different μ_p, the efficiency frontier for his portfolio can be constructed.

Figure 11 shows the efficient frontiers constructed using *RC* (black solid line), its overall bounds (upper bound: dashed line, right; lower bound: grey line, extreme left), as well as using its 95% confidence bands (dash-dotted lines) for the c-jpm asset pair in 2007. The plots are averages across the sample period of 251 days and are given in annualized percentages. The risk-return tradeoff is generally pessimistic due to the credit crisis which impacted both of these stocks negatively. The plots however shows the worst- and best-case scenarios to be more optimistic than implied by the confidence bands. It also shows that the efficiency frontier of the portfolios estimated using *RC* lies closer to the upper bound (worst-case scenario) than the lower bound (best-case scenario). While a portfolio manager might consider the 95% confidence band on the left to be overly conservative, he might deem the upper bound scenario to be realistic, worth planning for and informing his investors about.

Figure 11: Portfolio Efficiency Frontiers with Returns on the y-Axis and Standard Deviation on the x-Axis (Both Are Annualized in %)



Notes: The efficient frontiers are constructed using *RC* (black solid line, middle), upper (dashed line, right) and lower (grey line, extreme left) identification bounds and 95% confidence levels (dash-dotted lines). All plots are averages across the sample period of 251 days.

6. Conclusion

Estimating realized covariance and correlations is problematic due to data problems of asynchronous observations and microstructure noise. While different bias-correction methods exist, they often involve making assumptions about the latent noise process and quote/trade arrival-time process. This paper posits that rather than attempting to obtain point identification of the estimators that have to be bias corrected, a more robust approach can be used by way of partial identification (MANSKI, 1995).

We identify bounds due to the presence of asynchronicity by using the partial identification approach of HOROWITZ and MANSKI (2006) for missing data and the bounds due to microstructure noise by using the approach of HOROWITZ and MANSKI (1995) for treatment of contaminated and corrupted data to estimate the identification region. The bounds due to microstructure noise are estimated under two different error models: the contaminated sampling and the corrupted sampling schemes, where the contaminated sampling scheme makes the additional assumption that the error process is random and independent of the sampling process.

Our simulation study shows that bounds provide good coverage of the *RC* and *RCorr* estimators and their bias-corrected estimators via subsampling. Furthermore, we show how the tuning parameters in the estimation (namely the tolerance used to assign missingness and the assumed upper bound of noise levels) influence the widths of the estimated bounds.

We show via an empirical application that the overall bounds under the corrupted sampling scheme provide a high degree of coverage of the estimators and the subsampled estimators both in the pre-crisis and in the crisis period. However under the contaminated scheme, the coverage is unsatisfactory, which indicates that the noise process cannot be assumed to be independent of the efficient price process. This result is expected under market microstructure theory and in line with findings of Phillips and Yu (2006) and Barndorff-Nielsen et al. (2008). This gives indication that the estimated bounds are tight, and that the estimate of the upper bound of microstructure noise is reasonable.

We forecast the bounds of the realized correlations using the HAR model (Corsi, 2003) for 1-step and 10-step periods, and find that the forecasted bounds are tight with excellent coverage despite dealing with data during the volatile crisis period. While the accuracy of point estimators declines greatly under the 10-step forecast, the tightness and coverage of the bounds remain stable under multistep forecasting.

Applications of these bounds can be in financial risk management, such as for forecasting of the VaR, and in portfolio management, where best- and worst-case scenarios can be more reliably drawn when data problems are prevalent.

References

Aït-Sahalia, Y., P. A. 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, pp. 351–416.

Bandi, F. M., and J. R. Russell (2008), "Microstructure Noise, Realized Variance and Optimal Sampling", *The Review of Economic Studies*, 75, pp. 339–369.

Barndorff-Nielsen, O., P. Hansen, A. Lunde, and N. Shephard (2008), "Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise", *Econometrica*, 76(6), 1481–1536.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011), "Multivariate Realised Kernels: Consistent Positive Semidefinite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading", *Journal of Econometrics*, 162(2), pp. 149–169.

BARNDORFF-NIELSEN, O. E., and N. SHEPHARD (2002), "Econometric Analysis of Realized Volatility and its use in Estimating Stochastic Volatility Models", *Journal of the Royal Statistical Society, Series B*, 64, pp. 253–280.

BARNDORFF-NIELSEN, O. E., and N. SHEPHARD (2004), "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics", *Econometrica*, 72(3), pp. 885–925.

CHEN, C., and L.-M LIU (1993), "Joint Estimation of Model Parameters and Outlier Effects in Time Series", *Journal of the American Statistical Association*, 88, pp. 284–297.

ČÍŽEK, PAVEL, and WOLFGANG HÄRDLE (2006), "Robust Econometrics", Discussion paper, Humboldt-Universität zu Berlin.

CORSI, F. (2003), "Simple Long Memory Models of Realised Volatility", Manuscript, USI.

CORSI, F. (2009), "A Simple Approximate Long-Memory Model of Realized Volatility", *Journal of Financial Econometrics*, 7(2), pp. 174–196.

CORSI, FULVIO, and FRANCESCO AUDRINO (2007), "Realized Correlation Tick-by-Tick", Working paper, University of St Gallen.

DACOROGNA, M. M., R. GENCAY, U. MULLER, and R. B. OLSEN (2001), *An Introduction to High-Frequency Finance*, Academic Press, San Diego.

EPPS, T. W. (1979), "Comovements in Stock Prices in the Very Short Run", *Journal of the American Statistical Association*, 74, pp. 291–296.

GRIFFIN, JIM E., and ROEL C. A. OOMEN (2011), "Covariance Measurement in the Presence of Non-Synchronous Trading and Market Microstructure Noise", *Journal of Econometrics*, 160, pp. 56–68.

HAYASHI, T., and N. YOSHIDA (2005), "On Covariance Estimation of Nonsynchronously Observed Diffusion Processes", *Bernoulli*, 11, pp. 359–379.

HOROWITZ, J. L., and C. F. MANSKI (2006), "Identification and Estimation of Statistical Functionals Using Incomplete Data", *Journal of Econometrics*, 132, pp. 445–459.

HOROWITZ, J. L., and C. F. MANSKI (1995), "Identification and Robustness with Contaminated and Corrupted Data", *Econometrica*, 63, pp. 281–302.

HOROWITZ, J. L., and C. F. MANSKI (2000), "Nonparametric Analysis of Randomised Experiments with Missing Covariate and Outcome Data", *Journal of the American Statistical Association*, 95, pp. 77–84.

JACOD, J., Y. LI, P. A. MYKLAND, M. PODOLSKIJ, and M. VETTER (2009), "Microstructure Noise in the Continuous Case: The Pre-Averaging Approach", *Stochastic Processes and their Applications*, 119(7), pp. 2249–2276.

Kalnina, I., and O. Linton (2006), "Estimating Quadratic Variation Consistently in the Presence of Correlated Measurement Error", Working paper, Department of Economics, LSE.

Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.

Mykland, P. A., and L. Zhang (2006), "ANOVA for Diffusions and Ito Processes", *Annals of Statistics*, 34, pp. 1931–1963.

Nolte, I., and V. Voev (2009), "Least Squares Inference on Integrated Volatility and the Relationship between Efficient Prices and Noise", Working paper, Warwick Business School.

O'hara, M (1995), *Market Microstructure Theory*, Blackwell Publishers Ltd, Oxford.

Phillips, P. C. B., and J. Yu (2006), "Comment on 'Realized Variance and Market Microstructure Noise'", *Journal of Business and Economic Statistics*, 24, pp. 202–208.

Renault, Eric, and Bas J. M. Werker (2011), "Causality Effects in Return Volatility Measures with Random Times", *Journal of Econometrics*, 160(1), pp. 272–279.

Roll, R. (1984), "A Simple Measure of the Effective Bid-Ask Spread in an Efficient Market", *Journal of Finance*, 39, pp. 1127–1139.

Tsay, R. S., D. Pena, and A. E. Pankratz (2000), "Outliers in Multivariate Time Series", *Biometrika*, 87(4), pp. 789–804.

Voev, V., and A. Lunde (2007), "Integrated Covariance Estimation Using High-Frequency Data in the Presence of Noise", *Journal of Financial Econometrics*, 5, pp. 68–104.

Vortelinos, Dimitrios I. (2010), "The Properties of Realised Correlation: Evidence from the French, German and Greek Equity Markets", *The Quarterly Review of Economics and Finance*, 50(3), pp. 273–290.

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, pp. 1394–1411.

Zhang, Lan (2011), "Estimating Covariation: Epps Effect, Microstructure Noise", *Journal of Econometrics*, 160, pp. 33–47.

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

SUMMARY

This paper argues that the inherent data problems make precise point identification of realized correlation difficult but identification bounds in the spirit of MANSKI (1995) can be derived. These identification bounds allow for a more robust approach to inference especially when the realized correlation is used for estimating other risk measures. We forecast the identification bounds using the HAR model of CORSI (2003) using data during the year of onset of the credit crisis and find that the bounds provide good predictive coverage of the realized correlation for both 1- and 10-step forecasts even in volatile periods.