HIGH FREQUENCY ANALYSIS OF LEAD-LAG RELATIONSHIPS BETWEEN FINANCIAL MARKETS.

Frank de Jong
Theo Nijman
Tilburg University

February 1995

Mailing address:

Frank de Jong
Department of Econometrics
Tilburg University
PO BOX 90153
5000 LE TILBURG
The Netherlands

phone: +31-13-662911
fax: +31-13-663280
Email: F.deJong@KUB.NL

Thanks are due to Peter Bossaerts, Peter Schotman and Bas Werker for useful comments. Of course, the authors remain responsible for all errors.

Abstract

High frequency data are often observed at irregular intervals, which complicates the analysis of lead-lag relationships between financial markets. Frequently, estimators have been used that are based on observations at regular intervals, which are adapted to the irregular observations case by ignoring some observations and imputing others. In this paper we propose an estimator that avoids imputation and uses all available transactions to calculate (cross) covariances. This creates the possibility to analyze lead-lag relationships at arbitrarily high frequencies without additional imputation bias, as long as weak identifiability conditions are satisfied. We also provide an empirical application to the lead-lag relationship between the SP500 index and futures written on it.

1. Introduction.

Lead-lag relationships have been analyzed between many financial markets. A prime example is the link between the index futures and the cash market, where many researchers have found that the futures market leads the cash market (see e.g. Kawaller, Koch and Koch (1987), Stoll and Whaley (1990), Chan (1992) and Grünblicher, Longstaff and Schwartz (1994)). Others considered the relationship between the stock market and the option market. Stephan and Whaley (1990) find that the stock market leads the option market; this phenomenon is explained by Chan, Chung and Johnson (1993). Also, an increasing number of securities is traded on more than one financial market e.g. securities from many European countries outside the UK are traded on London's SEAQ International market in the domestic currency. Analysis of the lead-lag relationship between these markets would be yet another example.

In order to analyze information flows between markets on short time intervals, high frequency data are required. Typically, all transactions for some sample period are available for analysis. However, the statistical analysis of transactions data is often hampered by the fact that the clock time interval between such observations is varying. For some research questions, such as most micro-structure issues, the differences in clock time interval are not very important and one relies on estimating models in transaction time. However, for the analysis of information flows between markets the clock time is of utmost importance. The usual approach to tackle the problem of irregularly spaced observations is to split the time axis in fixed length intervals of, say, 5 minutes, and use the last observation recorded in that interval in the statistical analysis. This approach has two important drawbacks, however:

(i) If the intervals are small and trading is not very frequent, some intervals may contain no observation. This is referred to as the non-trading or non-synchronous trading problem. Another cause of missing observations are imperfections in data collection, e.g. errors on the data file, which sometimes cause a loss of observations. In both cases, ad hoc procedures to deal with missing observations must then be invoked.

(ii) On the other hand, in periods where trading is busy, a lot of observations are thrown away. This makes the statistical analysis less efficient. The loss of efficiency is an especially serious problem if busy trading is associated with large price changes, which is usually the case.

In this paper we propose an estimator that avoids arbitrary imputation methods. This creates the possibility to analyze lead-lag relationships at arbitrarily high frequencies without additional imputation bias, as long as weak identifiability conditions are satisfied.

The plan of the paper is as follows. In section 2 we introduce a consistent estimator of the covariances and correlations of interest from irregularly spaced data. In section 3 we derive the large sample distribution of these estimators. In section 4 we discuss some potential extensions of the method. Section 5 contains an empirical application to the lead-lag relationship between the S&P500 index and the futures on this index. Section 6 concludes. Technical details are discussed in the appendices.

2. Estimation of correlations in real time with irregularly spaced observations.

In this section, we present a method for estimating correlations between returns from irregularly spaced transactions data. The underlying model is a discrete

time process at an arbitrary time interval, not a continuous time process. We first consider the case where the returns have zero mean and there are no deterministic components in the model[1]. Let p_t and q_t denote the (logarithm of) levels of the two price series under consideration, where t is the clock-time index. The price *levels* are assumed to be non-stationary processes, which are stationary after differencing. Denote the cross-covariance function of the underlying *returns* (one-period price changes) by

$$(1) \qquad \gamma_k = \mathrm{Cov}(\Delta p_t, \Delta q_{t-k}), \qquad \Delta p_t \equiv p_t - p_{t-1}, \quad \Delta q_t \equiv q_t - q_{t-1}.$$

If the price levels were observed at every point, the covariances γ_k could be estimated efficiently by the usual expressions. However, when using transactions data there are potentially a lot of time intervals with no new observation on the price level. One way to 'solve' this problem is to impute a zero return for this interval, but that will bias the usual covariance estimators towards zero. In order to obtain an unbiased covariance estimator, we use the differences between observations on the price level over more than one interval. We then infer the covariances of the underlying but unobserved one-period returns from the cross-products of these more-period returns. In this section, we discuss this method in some detail.

We index the observations on p_t by the index i and the observations on q_t by the index j, and denote the total number of observations by N and M, respectively. The differences between two observed price levels can be expressed as sums of the returns of the unobserved underlying price process

$$(2) \qquad p_{t_{i+1}} - p_{t_i} = \sum_{t=t_i+1}^{t_{i+1}} \Delta p_t$$

where t_i denotes the clock-time index of the i^{th} observation. The cross product of price changes on the two markets can thus be written as

$$(3) \qquad y_{ij} \equiv (p_{t_{i+1}} - p_{t_i})(q_{t_{j+1}} - q_{t_j}) = \sum_{t=t_i+1}^{t_{i+1}} \Delta p_t \cdot \sum_{s=t_j+1}^{t_{j+1}} \Delta q_s .$$

The expectation of this cross-product is a linear combination of the cross-covariances γ_k of the underlying processes

$$(4) \qquad E(y_{ij}) = E\left(\sum_{t=t_i+1}^{t_{i+1}} \Delta p_t \cdot \sum_{s=t_j+1}^{t_{j+1}} \Delta q_s \right) = \sum_{t=t_i+1}^{t_{i+1}} \sum_{s=t_j+1}^{t_{j+1}} \gamma(t-s),$$

where the expectation in (4) is conditional on the observed transaction times $(t_i, t_j, t_{i+1}, t_{j+1})$. Let $x_{ij}(k)$ denote the number of times that $\gamma(k)$ appears in this expression. In appendix A the following expression for the $x_{ij}(k)$ is derived:

$$(5) \qquad x_{ij}(k) = \max(0, \min(t_{i+1}, t_{j+1} - k) - \max(t_i, t_j + k)).$$

An important property of the x_{ij}'s is that they are functions of the transaction times t_i only, not of the observed prices. Therefore, we replace the conditioning on the transaction times by a conditioning on the x_{ij}'s and write $E(y_{ij})$ as a linear combination of the covariances $\gamma(k)$, k=-K,..,K, as follows

$$(6) \qquad E(y_{ij} \mid x_{ij}) = \sum_{k=-K}^{K} x_{ij}(k)\gamma(k).$$

Our estimation method is based on the fact that equation (6) can be considered as a regression equation with the unknown cross-covariances γ_k as parameters and

the coefficients x_{ij} as explanatory variables. In vector notation, the regression equation reads

(7) $$y_{ij} \equiv x'_{ij}\gamma + e_{ij}$$

The covariances can then be estimated by ordinary least squares on the observations of y_{ij} and the constructed x_{ij}'s [2]. In principle, all possible differences between observed prices can be used to construct an x_{ij} and y_{ij}. However, we can confine ourselves to differences of *adjacent* observations. The reason for this is that differences of non-adjacent observations always can be written as *exact* linear combinations of differences of adjacent observations. For example, consider

(8) $$(p_{t_{i+2}}-p_{t_i})(q_{t_{j+1}}-q_{t_j}) =$$

$$(p_{t_{i+2}}-p_{t_{i+1}})(q_{t_{j+1}}-q_{t_j}) + (p_{t_{i+1}}-p_{t_i})(q_{t_{j+1}}-q_{t_j}) = y_{i+1,j} + y_{ij}$$

For this reason, non-adjacent observations do not add information and can be omitted. All in all, N times M cross-products y_{ij} are available for the analysis. It is not necessary to use all of them, however, if the number of non-zero cross-covariances is limited, say to K. In that case, all cross products where $|t_{i+1}-t_j| \geq K$ and $|t_i-t_{j+1}| \geq K$ can be omitted because there will be no non-zero elements in x_{ij} in that case. Let all useful observations be contained in the design matrix X and vector of observations y as follows

$$(9) \qquad X = \begin{pmatrix} x_{11}(-K) & .. & x_{11}(K) \\ x_{12}(-K) & .. & x_{12}(K) \\ . & & . \\ x_{1M}(-K) & & x_{1M}(K) \\ . & & . \\ x_{NM}(-K) & .. & x_{NM}(K) \end{pmatrix}, \qquad y = \begin{pmatrix} y_{11} \\ y_{12} \\ . \\ y_{1M} \\ . \\ y_{NM} \end{pmatrix}$$

Following Cohen et al. (1983) and Lo and MacKinlay (1990,1991), we assume that the order arrival process is independent of the price process. Under this assumption, and if $X'X$ is invertible and weak regularity conditions are satisfied, the OLS estimator $\hat{\gamma} \equiv (X'X)^{-1}X'y$ is a consistent estimator for the unconditional covariances of $\gamma = (\gamma_{-K},..,\gamma_K)'$. A necessary condition for consistency is that all omitted covariances (i.e. of order > K) are indeed equal to zero. If these covariances are not equal to zero the regression model will suffer from an omitted variables bias. Hence, even if one wants to estimate, say, only the first order correlation, one should estimate the whole vector of non-zero covariances.

The proposed estimator is more general than the models proposed by Cohen et al. (1983) and Lo and MacKinlay (1991), because we do not assume a particular process for the order arrival. As long as the order arrival process is exogenous to the price changes our estimator yields consistent estimates of the covariances in clock-time.

We shall now discuss some special cases of our estimator.

Example 1. Prices observed in every period.

The first case we discuss is the standard case where p_t and q_t are observed in every period. In this case, only the usual first differences Δp_t and Δq_t need to be considered. When estimating cross-covariances -K to K, the design matrix becomes

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ . & & & & . \\ . & & & & . \\ 0 & . & . & 0 & 1 \end{pmatrix}, \qquad y = \begin{pmatrix} y_{1,1-K} \\ y_{1,2-K} \\ . \\ y_{1,1+K} \\ y_{2,2-K} \\ . \\ . \\ y_{N,N+K} \end{pmatrix} = \begin{pmatrix} \Delta p_1 \Delta q_{1-K} \\ \Delta p_1 \Delta q_{2-K} \\ . \\ \Delta p_1 \Delta q_{1+K} \\ \Delta p_2 \Delta q_{2-K} \\ . \\ . \\ \Delta p_N \Delta q_{N+K} \end{pmatrix}$$

Obviously, the $X'X$ matrix is a diagonal matrix $N \cdot I_{2K+1}$, and $X'y$ is a vector with typical elements $\sum \Delta p_t \Delta q_{t+k}$, so that the OLS estimator is equal to the usual covariance estimator, $\hat{\gamma}_k = N^{-1} \sum \Delta p_t \Delta q_{t+k}$.

Example 2. Regularly missing observations.

Now suppose that the prices are not observed in every period, but on regularly spaced intervals. To choose the simplest example, suppose that p_t and q_t are observed every second period. The useful cross-products then are

$$y_{ij} = (p_t - p_{t-2})(q_{t-k} - q_{t-k-2}) = (\Delta p_t + \Delta p_{t-1})(\Delta q_{t-k} + \Delta q_{t-k-1}) \quad \Rightarrow$$

$$E(y_{ij}) = \gamma_k + 2\gamma_{k-1} + \gamma_{k-2}$$

The design matrix X in the simplest case that K=1 takes the form

$$X = \begin{pmatrix} 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ . & & & & . \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 \end{pmatrix}$$

It is clear that the first and second column are exact multiples of each other, so that there is extreme multicollinearity. Therefore X′X is singular and not all cross-covariances can be estimated. This argument can be generalized to the statement that either complete observations or some *irregularly* missing observations are necessary to estimate all covariances. Throughout this paper we shall assume that such identifying conditions are satisfied.

From the estimates of the autocovariances, estimates of the autocorrelations can be computed in the usual way. The cross-correlation function is defined as the cross-covariances, scaled by the square root of the product of the estimated variances of Δp_t and Δq_t

$$(10) \qquad \hat{\rho}(k) = \frac{\hat{\gamma}(k)}{[\hat{\gamma}_p(0)\hat{\gamma}_q(0)]^{1/2}}.$$

3. Large sample distribution of the estimators.

In this section we derive the large sample distribution of the estimators derived in the previous section. This large sample distribution can be used to test for the significance of lead-lag effects. We start from the usual result that the regression estimator is asymptotically normal and that its variance-covariance matrix can be expressed as

$$(11) \qquad \Omega = (X'X)^{-1}X'E(ee')X(X'X)^{-1}.$$

Two estimators of Ω will be considered. Under strong additional assumptions it is possible to obtain an analytic expression for $\Sigma=E(ee')$. Subsequently we present a more robust, White-type estimator of Ω.

In order to derive the first estimator of Ω, assume that Δp_t and Δq_t are generated by the same innovations ε_t, but with different MA coefficients

$$\Delta p_t = \phi_0\varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + = \Sigma_{i=0}^{\infty}\phi_i\varepsilon_{t-i}$$

(12)

$$\Delta q_t = \theta_0\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + = \Sigma_{i=0}^{\infty}\theta_i\varepsilon_{t-i}$$

Note that this assumption implies that the level variables p_t and q_t are cointegrated[3]. In Appendix B it is shown that the elements of Σ can be expressed as

(13) $$\sigma_{ij,gh} = x'_{ig}\gamma_p \cdot x'_{jh}\gamma_q + x'_{ih}\gamma \cdot x'_{jg}\gamma + (\mu_4-3\sigma^4)f(\theta,\phi)$$

where γ_p and γ_q denote the auto-covariances of $\{p_t\}$ and $\{q_t\}$, respectively, μ_4 the fourth moment of the innovations, and $f(\theta,\phi)$ is an expression in the MA coefficients. If the errors are non-normal, the MA coefficients and the fourth moment of the innovation have to be calculated in order to estimate the $(\mu_4-3\sigma^4)f(\theta,\phi)$ term. This makes empirical application of this result cumbersome.

An alternative and more robust way to calculate standard errors is a White (1980) type estimator, where the expectation of $X'ee'X$ is estimated by a summation over all observations for which $E(e_{ij}e_{gh})$ is non-zero. Thus, the estimator of Ω becomes

(14) $$\Omega = (X'X)^{-1}(\Sigma_{ij}\Sigma_{gh}x_{ij}x'_{gh}e_{ij}e_{gh}\cdot I(\sigma_{ij,gh}\neq 0))(X'X)^{-1},$$

where I(.) is an indicator function which equals one if $\sigma_{ij,gh} \neq 0$, and zero elsewhere. The latter property is easily checked from equation (13). This estimator will be consistent for Ω under much weaker assumptions on the data generating process. For example, we do not need normality, nor the restrictive assumption that both series (p and q) are generated by the same innovations.

Note that the number of non-zero covariances used to calculate the standard errors in (13) or to calculate the indicator in (14) can be smaller than the number of covariances actually estimated. For example, we can estimate 10 covariances, but calculate the standard errors under the hypothesis that all but the first are zero. This will simplify and speed up the calculations of the standard errors considerably.

4. Extensions of the method.

In this section, we discuss two potential extensions of our method to estimate covariances on irregularly spaced data. The first extension is the inclusion of a latent bid-ask spread, which is very relevant for the applications to financial time series. The second extension is the inclusion of additional observed explanatory variables.

To start with the first extension, suppose that the observed prices can be decomposed in an equilibrium price π_i plus or minus a fixed bid-ask spread, $\delta=S/2$. We are interested in estimating the autocorrelations of $\Delta\pi_t$. Define a binomial indicator b_i, which can take values +1 and -1, such that

(15) $$p_i = \pi_i + b_i\delta,$$

so that the observed price differences can be written as

(16) $$p_{t_{i+1}}-p_{t_i} = \pi_{t_{i+1}}-\pi_{t_i} + (b_{i+1}-b_i)\delta = \sum_{t=t_i+1}^{t_{i+1}} \Delta\pi_t + (b_{i+1}-b_i)\delta.$$

First, consider the case where we do not know whether the transaction is at the bid or at the ask, hence b_i is unobserved. We now introduce some strong assumptions on the bid-ask indicator: b_i has expectation zero and is uncorrelated with both its own past and with the price and transaction time processes. Note that these are basically Roll's (1984) assumptions, and therefore our method can be seen as an adaption of Roll's estimator. Under these assumptions, the expectation of the cross-product of price differences is

(17) $$E[(p_{t_{i+1}}-p_{t_i})(p_{t_{j+1}}-p_{t_j})] = x'_{ij}\gamma + E[(b_{i+1}-b_i)(b_{j+1}-b_j)]\delta^2 = x'_{ij}\gamma + d_{ij}\delta^2$$

where γ now is the vector of covariances of the equilibrium price changes $\Delta\pi_t$, and the the new regressor d_{ij} is defined as follows: $d_{ij}=2$ if i=j, $d_{ij}=-1$ if j=i+1 or j=i-1, and $d_{ij}=0$ otherwise. Note that the values of d_{ij} do not depend on the time of the transactions, only on the sequencing.

Equation (17) is a straightforward extension of the original model (5), and the estimators and standard errors described in the previous sections can be applied to this model immediately. Twice the square root of the estimated coefficient of d_{ij} can be used as an estimator for the realized bid-ask spread. This estimator of the bid-ask spread is similar in spirit to the one proposed by Roll (1984) and Richardson and Smith (1991), who use a GMM estimator to estimate the mean, variance and bid-ask spread on series of overlapping returns. Our estimator is more general than Roll's estimator and Richardson and Smith's estimator because it allows for serial correlation in the equilibrium price process and for irregular trading intervals. However, the spread estimator suffers from the same weaknesses as Roll's estimator: it needs the assumption

that the bid-ask bounce is independent of the price process. Market microstructure theory suggests that this is a very unrealistic assumption. For example, in the Glosten-Milgrom (1985) model with only asymmetric information there is a bid-ask spread, but the serial correlation in observed prices is zero, hence Roll's and our estimator will estimate a zero spread.

The second extension is the inclusion of observed regressors other than the x_{ij}'s. Conceptually, this is trivial as it extends the model to

$$(18) \qquad y_{ij} = x'_{ij}\gamma + z'_{ij}\beta + e_{ij}.$$

As long as the z_{ij}'s are uncorrelated with the error term, nothing changes and the OLS estimators and the robust standard errors will be consistent. This extension is useful if the bid-ask indicator b_i is observed. In that case, the observed cross-products $(b_{i+1}-b_i)(b_{j+1}-b_j)$ can be added to the model as additional regressors:

$$(19) \qquad E(y_{ij}) = x_{ij}'\gamma + (b_{i+1}-b_i)(b_{j+1}-b_j)\cdot\delta^2 + e_{ij}.$$

In this case, there will be no bias in the effective spread estimates even if the b_i series is serially correlated or depends on previous price changes.

5. Empirical application.

In this section we present an empirical application of the proposed estimator to the lead-lag relationship between the S&P 500 stock index and futures on this index. As stated in the introduction, this is a well-studied relationship, with the general conclusion that the futures market leads the cash market. Typically, researchers have used five minute intervals, where few observations are missing.

In this section, we also present results at the one minute interval, at which more intervals without trade occur in the futures market. Since the stock market index is adjusted every minute there are no missing data points on the index unless the frequency at which the data are analyzed is even higher than one minute.

Following Stoll and Whaley (1990), the relation between cash index prices and futures prices can be expressed simply as

$$F_t = S_t \exp[(r-d)(T-t)], \tag{20}$$

where F_t denotes the futures price, S_t the cash price, (r-d) the interest rate minus the convenience yield (dividends), assumed constant, and T the expiration date of the futures contract. From (20) it is easily seen that there is an exact theoretical relation between the logarithmic returns on the cash index and the futures:

$$R_t^F = (r-d) + R_t^S. \tag{21}$$

In practice, the equality does not always hold exactly. An obvious cause of these deviations are measurement errors and the effect of the bid-ask spread. Another explanation, which for the purpose of our paper is more interesting, is given by potential differences in the speed at which information is disseminated to both markets or the limited ability of index arbitrage, which involves trading in a large number of assets. Therefore, it is interesting to assess whether the returns on one market are predictable from the returns in the other market.

Stoll and Whaley (1990) investigate this question for the US indexes. Stoll and Whaley use observations on all transactions or quote changes of the

S&P 500 index and the Major Market Index (MMI) and the futures on these indices. The trading day is divided into intervals of 5 minutes. The first prices to be observed in these intervals are then used to construct 5-minute returns in both the cash index and futures markets. This creates some problems if there are no transactions in some interval. Usually, a zero return for these periods is imposed. Stoll and Whaley's empirical methodology is in two steps. First, they calculate the auto- and cross-correlations of R^S and R^F. The SP500 cash index returns show strong positive serial correlation. The futures returns are almost serially uncorrelated. Individual stock returns tend to be negatively serially correlated due to the bid-ask bounce.

These results are exactly in the direction predicted by Lo and MacKinlay (1991), who show that the returns of a continuously trading market must lead the observed returns from a market with a positive probability of non-trading. However, the magnitude of the correlations found by Stoll and Whaley cannot be explained by the actually observed probability of non-trading. Chan (1992) corroborates these conclusions on the Major Market Index, which consists of 20 large stocks and is therefore less prone to non-trading problems. The futures returns lead the MMI index return by 15 minutes and also tend to lead individual stock returns. Especially market-wide information seems to be processed faster in the futures market.

The conclusion of the literature therefore is that the futures market processes new information faster than the cash index market. In this paper we shall investigate this proposition using the covariance estimators developed in the previous sections. The estimator deals naturally with intervals without new observations on the index or futures price. Therefore, the analysis can be performed on a higher frequency than the usual 5 minutes[4].

Our data concern spot and futures prices of the S&P 500 index, obtained from the ISSM. The sample is from the last quarter of 1993 [5]. The index prices

are time stamped exactly at the full minute, whereas the timing of the futures prices is exact up to one second. The data are discretized by taking the last trade or index report in a given interval as the value of the level variable for that interval. If there is no single trade in an interval, this observation is missing. We consider observations on the futures that expire in December 1993 (before 15/12) and March 1994 (after 15/12). As usual when dealing with intra-day data, we exclude overnight returns from the analysis, as these cannot be expected to have the same covariance structure as within-day returns, see French and Roll (1986). We have nearly complete observations for the index. However, for the futures there are intervals without transactions. For example, at the one minute frequency, 13% of the intervals does not contain a new observation.

As a first step in the analysis, we estimate the autocorrelations of the futures price changes and the index changes. Table 1 reports the autocorrelation estimates of the index and Table 2 those of the futures returns. We consider time intervals of ten and five minutes, as well as a one minute interval. In all empirical results, the variance-covariance matrix of the estimates is calculated under the assumption that only the variance and the first covariances of the returns are non-zero. First, we consider the results on a five and ten minute interval. Following Chan (1992), the maximum order of correlation considered is six. The index returns show little serial correlation on a ten minute interval, and positive first order correlation on a five minute interval, but further lags are not significant. The futures returns are serially uncorrelated at both the five and ten minute interval. If we increase the frequency of observation to one minute, a different pattern emerges. For the index, the serial correlations are significantly positive, up to order eight. The estimated autocorrelations are smaller than the estimates in Harris et al. (1994), probably as a result of the different sample period used. The first order autocorrelation in the futures returns is significantly negative. This is very likely due to the bid-ask bounce

of the futures contract. There is no significant higher order serial correlation in the futures returns, which shows that all relevant information is immediately reflected in the futures prices, even on such a high frequency as one minute.

We now turn to the lead-lag structure of cash and futures price changes. The cross-correlations between futures and index returns are reported in Table 3. These are defined as the cross-covariances, $Cov(R^x_t, R^f_{t-k})$, divided by the standard deviation of the index and futures return on the same interval. A positive correlation for $k>0$ indicates that the futures returns have predictive ability for the index returns. The results of this table are unambiguous: at all intervals, the futures returns significantly lead the index returns. The time span of this correlation is at least ten minutes, given the significant first order cross-correlation at the ten minute interval. At the one minute frequency, up to ten lead correlations of the futures are significant. This conclusion is confirmed by the joint significance tests of all lead coefficients in Table 4. On the other hand, there is no evidence that the index returns lead the futures returns by more than five minutes, because the cross-correlations for $k<0$ are insignificant at the five and ten minute intervals. At the one minute interval, there is some lead correlation from the index to the futures returns, but only up to two minutes.

The cross-correlations are stronger than is predicted by the autocovariances in the index alone (cf. Boudoukh, Richardson and Whitelaw (1994)). Hence, the correlation cannot be due solely to thin and nonsychronous trading in the index alone. An alternative explanation, put forward by Chan (1993) and Bossaerts (1993) is based on differential information in markets. If firm specific information cannot be separated from market wide information in the individual stock markets, index returns will be positively serially correlated, despite the fact that the individual stock returns are serially

uncorrelated. If the futures market reflects only market wide information it will lead the returns on the cash index.

6. Conclusions.

In this paper we have developed a method for estimating covariances of non-stationary time series with irregularly spaced observations. Under weak conditions, this estimator is consistent under any pattern of missing observations. Several extensions to include latent or deterministic variables are developed.

We apply the method to the lead-lag relation between stock market index returns and index futures returns. An analysis on a one minute frequency reveals that the futures lead the cash index by at least ten minutes, whereas the cash index leads the futures by at most two minutes. Another application of our estimator can be found in De Jong, Mahieu and Schotman (1995). In that paper, we apply the proposed methods to exchange rates. In particular, we study lead-lag patterns between the actual Yen/Dmark exchange rate and the exchange rate implied by cross-arbitrage via the US dollar exchange rates.

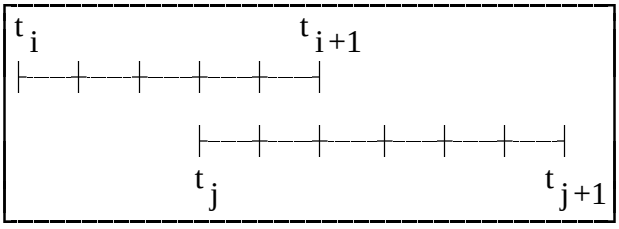
Appendix A. An expression for x_{ij}.

Recall the definition of x_{ij} in (4). In this appendix we show how to simplify the calculations necessary to obtain the elements of x_{ij}. By changing the index of summation from i-j to k and working out the resulting expression we obtain

$$x'_{ij}\gamma \equiv \sum_{t=t_i+1}^{t_{i+1}} \sum_{s=t_j+1}^{t_{j+1}} \gamma(t-s) = \sum_{t=t_i+1}^{t_{i+1}} \sum_{k=t-t_{j+1}}^{t-t_j-1} \gamma(k) = \sum_{k=t_i-t_{j+1}+1}^{t_{i+1}-t_j-1} x_{ij}(k)\gamma(k),$$

What remains to be determined is the coefficient $x_{ij}(k)$ of $\gamma(k)$. To facilitate the derivation of this number, in Figure A the intervals $[t_i,t_{i+1}]$ and $[t_j,t_{j+1}]$ are graphed.

Figure A. Overlapping intervals between two pairs of observations.



The number of correlations $\gamma(k)$ between the price changes over these intervals can be determined by shifting the $[t_j,t_{j+1}]$ interval by k periods to the right, to obtain $[t_j+k,t_{j+1}+k]$. The coefficient of $\gamma(k)$ is exactly equal to the number of periods in the overlap of the intervals $[t_i,t_{i+1}]$ and $[t_j+k,t_{j+1}+k]$. If the set of overlapping periods is not empty, the time index of the upper bound of the overlapping interval is $\min(t_{i+1},t_{j+1}+k)$, and the time index of the lower bound of the overlapping interval is $\max(t_i,t_j+k))$. The number of covariances $\gamma(k)$ is thus equal to the difference between the upper and lower bounds of this interval. If the intervals do not overlap, $\gamma(k)$ is by definition equal to 0. The upshot of this analysis is the following expression

$$x_{ij}(k) = \max(0, \min(t_{i+1}, t_{j+1}+k) - \max(t_i, t_j+k)).$$

If the maximal order of correlation is restricted a priori, so that $\gamma(k)=0$ for $|k|>K$, then the summation over k is truncated between -K and K, as follows

$$x'_{ij}\gamma = \sum_{k=-K}^{K} x_{ij}(k)\gamma(k),$$

where the definition of $x_{ij}(k)$ remains unchanged. Using this expression for x_{ij} reduces the computation time substantially because double summations are avoided.

In the case of estimating auto-covariances, the coefficients $x_{ij}(-k)$ should be added to $x_{ij}(k)$ for all k=1,..,K. Note that $x_{ij}(0)$ is not changed. The dimension of the regression model is thus reduced to K+1.

Appendix B. The covariance structure of the error terms.

Let Δp_t and Δq_t have the following Wold representations, driven by the same innovations ε_t but with different MA parameters $\{\phi_i\}$ and $\{\vartheta_i\}$

$$\Delta p_t = \sum_{i=0}^{K} \phi_i \varepsilon_{t-i}$$

$$\Delta q_t = \sum_{i=0}^{K} \vartheta_i \varepsilon_{t-i}$$

The error terms of the regression equation (5) are

$$e_{ij} = y_{ij} - E(y_{ij}) = \Big(\sum_{t=t_i+1}^{t_{i+1}} \Delta p_t\Big)\Big(\sum_{s=t_j+1}^{t_{j+1}} \Delta q_s\Big) - \sum_{t=t_i+1}^{t_{i+1}} \sum_{s=t_j+1}^{t_{j+1}} \gamma(t-s)$$

The covariance between two such errors is

$$E(e_{ij}e_{gh}) = E\left[\Big(\sum_{t=t_i+1}^{t_{i+1}} \Delta p_t\Big)\Big(\sum_{s=t_j+1}^{t_{j+1}} \Delta q_s\Big)\Big(\sum_{u=t_g+1}^{t_{g+1}} \Delta p_u\Big)\Big(\sum_{v=t_h+1}^{t_{h+1}} \Delta q_v\Big)\right] -$$

$$\left(\sum_{t=t_i+1}^{t_{i+1}} \sum_{s=t_j+1}^{t_{j+1}} \gamma(t-s)\right)\left(\sum_{u=t_g+1}^{t_{g+1}} \sum_{v=t_h+1}^{t_{h+1}} \gamma(u-v)\right)$$

$$= \sum_{t=t_i+1}^{t_{i+1}} \sum_{s=t_j+1}^{t_{j+1}} \sum_{u=t_g+1}^{t_{g+1}} \sum_{v=t_h+1}^{t_{h+1}} \Big(E(\Delta p_t \Delta q_s \Delta p_u \Delta q_v) - \gamma(t-s)\gamma(u-v)\Big)$$

By application of the expression given in Brockwell and Davis (1987, p.220), for the expectation of the four-fold product $\Delta p_t \Delta q_s \Delta p_u \Delta q_v$ we obtain

$$E(e_{ij}e_{gh}) = \sum_{t=t_i+1}^{t_{i+1}} \sum_{s=t_j+1}^{t_{j+1}} \sum_{u=t_g+1}^{t_{g+1}} \sum_{v=t_h+1}^{t_{h+1}} \Big(\gamma_p(t-u)\gamma_q(s-v) + \gamma(t-v)\gamma(u-s) + (\mu_4-3\sigma^4) \sum_{i=0}^{K} \vartheta_i \vartheta_{i+s-t} \phi_{i+u-t} \phi_{i+v-t} \Big)$$

where γ_p and γ_q denote the auto-covariances of Δp and Δq, respectively, and σ^2 and μ_4 denote the second and fourth moment of the innovations ε_t [6].

The expression for the covariance considerably simplifies if the innovations ε_t are normally distributed. In that case, the $(\mu_4-3\sigma^4)$ term vanishes and the resulting expression contains only auto- and cross covariances and the fourfold summation can be split into products of double summations

$$E(e_{ij}e_{gh}) = \sum_{t=t_i+1}^{t_{i+1}} \sum_{s=t_j+1}^{t_{j+1}} \sum_{u=t_g+1}^{t_{g+1}} \sum_{v=t_h+1}^{t_{h+1}} \Big(\gamma_p(t-u)\gamma_q(s-v) + \gamma(t-v)\gamma(u-s) \Big) =$$

$$\Big(\sum_{t=t_i+1}^{t_{i+1}} \sum_{u=t_g+1}^{t_{g+1}} \gamma_p(t-u) \Big) \Big(\sum_{s=t_j+1}^{t_{j+1}} \sum_{v=t_h+1}^{t_{h+1}} \gamma_q(s-v) \Big) +$$

$$\Big(\sum_{t=t_i+1}^{t_{i+1}} \sum_{v=t_h+1}^{t_{h+1}} \gamma(t-v) \Big) \Big(\sum_{u=t_g+1}^{t_{g+1}} \sum_{s=t_j+1}^{t_{j+1}} \gamma(u-s) \Big)$$

In shorthand, using the definition of x_{ij}, this can be written as (13).

References.

Bossaerts, Peter (1993), "Transaction prices when insiders trade portfolios", *Finance* **14**, 43-60.

Boudoukh, Jacob, Matthew Richardson and Robert Whitelaw (1994), "A tale of three schools: Insights on autocorrelations of short-horizon stock returns", *Review of Financial Studies* **7**, 539-573.

Chan, Kalok (1992), "A further analysis of the lead-lag relationship between the cash market and the stock index futures market", *Review of Financial Studies* **5**, 123-152.

Chan, Kalok (1993), "Imperfect information and cross-autocorrelation among stock prices", *Journal of Finance* **48**, 1211-1230.

Chan, K., Y.P. Chung and H. Johnson (1993), "Why option prices lag stock prices: a trading based explanation", *Journal of Finance* **48**, 1957-1967.

Cohen, K., G. Hawawimi, S. Maier, R. Schwartz and D. Whitcomb (1983), "Friction in the trading process and the estimation of systematic risk", *Journal of Financial Economics* **12**, 263-278.

De Jong, Frank, Ronald Mahieu and Peter Schotman (1995), "The dynamics of the actual and dollar implied Yen/Dmark cross exchange rate", in preparation.

French, K. and R. Roll (1986), "Stock return variances: The arrival of information and the reaction of traders", *Journal of Financial Economics* **17**, 5-26.

Grünblicher, A., F. Longstaff and E. Schwartz (1994), "Electronic screen trading and the transmission of information: An empirical examination", *Journal of Financial Intermediation* **3**, 166-187.

Hannan, E.J. (1960), *Time Series Analysis*, Methuen, London.

Harris, Lawrence, George Sofianos and James Shapiro (1994), "Program trading and intraday volatility", *Review of Financial Studies* **7**, 653-685.

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

Lo,Andrew and A. Craig MacKinlay (1990), "When are contrarian profits due to stock market overreaction?", *Review of Financial Studies*, **3**, 175-205.

Lo,Andrew and A. Craig MacKinlay (1991), "An econometric analysis of infrequent trading", *Journal of Econometrics*, **45**, 181-211.

Miller, Merton, Jayaram Muthuswamy and Robert Whaley (1994), "Mean reversion of Standard and Poor's 500 index basis changes: Arbitrage induced or statistical illusion?", *Journal of Finance* **49** (2), 479-513.

Richardson, Matthew and Tom Smith (1991), "Tests of financial models in the presence of overlapping observations", *Review of Financial Studies*, **4**, 227-254.

Roll, Richard (1984), "A simple implicit measure of the effective bid-ask spread in an efficient market", *Journal of Finance* **39**, 1127-1139.

Stephan, Jens and Robert Whaley (1990), "Intraday price change and trading volume relations in the stock and stock option markets", *Journal of Finance* **45**, 191-220.

Stoll, Hans and Robert Whaley (1990), "The dynamics of stock index and stock index futures returns", *Journal of Financial and Quantitative Analysis* **25**, 441-468.

White, Hal (1990), "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", *Econometrica* **48**, 817-838.

Table 1. Autocorrelations of index returns.

lag	10 minutes		5 minutes		1 minute	
0	**0.003869**	(7.53)	**0.001449**	(9.03)	**0.000166**	(19.83)
1	0.083	(1.37)	0.278*	(5.09)	0.195*	(5.84)
2	-0.023	(0.50)	0.037	(0.82)	0.176*	(7.08)
3	0.008	(0.19)	-0.023	(0.50)	0.144*	(7.16)
4	-0.032	(0.51)	-0.014	(0.31)	0.125*	(7.44)
5	0.020	(0.44)	0.007	(0.14)	0.094*	(5.87)
6	0.038	(0.59)	-0.011	(0.23)	0.082*	(4.66)
7					0.040*	(2.23)
8					0.052*	(2.97)
9					0.019	(0.74)
10					0.011	(0.44)
11					-0.005	(0.19)
12					0.005	(0.22)
13					0.009	(0.32)
14					-0.007	(0.23)
15					-0.006	(0.16)
nobs	823		1619		7989	
%missing	(0)		(0)		(0)	

Note: lag 0 denotes the variance of the series, other numbers are correlations. The numbers in parentheses are heteroskedasticity and serial correlation consistent t-statistics (calculated with one lag and lead window).

Table 2. Autocorrelations of futures returns.

lag	10 minutes		5 minutes		1 minute	
0	**0.004646**	(8.68)	**0.002179**	(12.57)	**0.000464**	(29.45)
1	-0.005	(0.08)	0.039	(0.86)	-0.287*	(13.39)
2	0.016	(0.26)	0.023	(0.49)	-0.028	(1.52)
3	0.010	(0.11)	-0.019	(0.32)	0.005	(0.29)
4	0.000	(0.00)	-0.004	(0.05)	0.012	(0.58)
5	0.019	(0.15)	0.024	(0.27)	-0.011	(0.41)
6	0.046	(0.32)	-0.047	(0.52)	-0.007	(0.24)
7					0.027	(0.61)
8					0.003	(0.06)
9					-0.013	(0.29)
10					0.011	(0.21)
11					0.037	(0.73)
12					-0.024	(0.53)
13					-0.023	(0.49)
14					0.002	(0.05)
15					0.026	(0.47)
nobs	760		1494		6807	
%missing	(0)		(1)		(14)	

Notes: see table 1.

Table 3. Correlations between index and future returns.

lag	10 minutes		5 minutes		1 minute	
-15					-0.005	(0.25)
-14					0.006	(0.36)
-13					0.013	(0.94)
-12					-0.031	(2.40)
-11					0.008	(0.60)
-10					0.001	(0.04)
-9					0.002	(0.17)
-8					0.012	(0.83)
-7					-0.002	(0.14)
-6	0.061	(1.06)	-0.010	(0.25)	0.014	(0.76)
-5	0.057	(1.30)	-0.018	(0.49)	-0.014	(0.87)
-4	-0.039	(0.57)	0.020	(0.69)	0.015	(0.98)
-3	0.025	(0.48)	-0.014	(0.29)	0.008	(0.47)
-2	-0.004	(0.10)	-0.002	(0.04)	0.033*	(3.15)
-1	0.008	(0.12)	0.075	(1.46)	0.164*	(9.19)
0	0.647*	(6.09)	0.514*	(7.43)	0.101*	(5.07)
1	0.311*	(4.43)	0.440*	(6.82)	0.171*	(7.23)
2	0.022	(0.50)	0.146*	(3.17)	0.168*	(7.11)
3	0.015	(0.33)	0.044	(1.25)	0.145*	(7.81)
4	-0.005	(0.08)	0.009	(0.28)	0.110*	(6.35)
5	0.006	(0.12)	0.003	(0.08)	0.103*	(6.37)
6	0.013	(0.41)	-0.014	(0.38)	0.058*	(4.32)
7					0.056*	(3.57)
8					0.023	(1.44)
9					0.056*	(3.83)
10					0.020	(1.25)
11					0.039*	(2.42)
12					0.025	(1.57)
13					0.010	(0.66)
14					0.032*	(2.41)
15					-0.001	(0.08)

Note: the entries in this table are estimates of the cross-correlations, i.e. $\mathrm{Cov}(\Delta s_t, \Delta f_{t-k})$ divided by the standard deviation of Δs_t and Δf_t. The numbers in parentheses are heteroskedasticity consistent t-statistics.

Table 4. Joint significance of 6 lead or lag covariances.

	10 minutes	5 minutes	1 minute
lag	1.01	6.73	99.35
lead	20.45	61.01	167.22

Notes: the entries are Wald (F-)statistics for the joint hypothesis that the lag (k<0) or lead (k>0) covariances are all equal to zero. The asymptotic distribution of this statistic is $\chi^2(6)$.

---

1 These will be introduced in the model in section 4.

2 In order to calculate *auto*-covariances of a time series with irregularly spaced observations, we have to change the definition of x_{ij} slightly because in that case $\gamma(k)=\gamma(-k)$.

3 For the empirical example in section 5, where we estimate cross-correlations between a stock index and index futures, this is a reasonable assumption.

4 The only study (to our best knowledge) which uses one minute returns is Harris et al. (1994). However, they calculate only autocorrelations and no lead-lag correlations between index and futures returns.

5 Not all trading days were reported on the tape. In total, we have only 19 complete trading days available. The maximum number of observations for the index series and the futures series are different because the trading day for futures is usually shorter than the period for which the index is reported.

6 This result corresponds to that found in Hannan (1960, p.39).