On the estimation of correlations for irregularly spaced time series

Jonas Andersson
Department of Finance and Management Science
Norwegian School of Economics and Business Administration
Helleveien 30
NO-5045 Bergen
Norway

June 26, 2007

Abstract

In this paper, the problem of calculating covariances and correlations between time series which are observed irregularly and at different points in time, is treated. The problem of dependence between the time stamp process and the return process is especially highlighted and the solution to this problem for a special case is given. Furthermore, estimators based on different interpolation methods are investigated. The covariances are in turn used to estimate a simple regression on such data. In particular, the difference of first order integrated processes, I(1) processes, are considered. These methods are relevant for stock returns and consequently of importance in e.g. portfolio optimization.

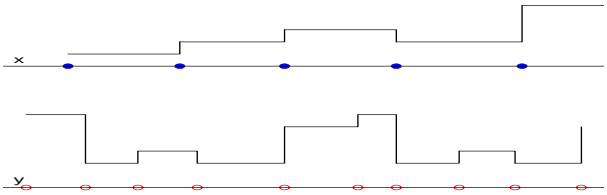
Keywords: Irregularly spaced time series, covariance, correlation, financial returns

1 Introduction

In many application of time series analysis, the problem of irregularly spaced observations has to be dealt with. A presently very interesting example is the increased use of intraday data from financial markets (e.g. Dacorogna et al., 2001; Campbell et al., 1997). Financial assets are certainly not traded at any predefined time points and neither are different assets traded at the same time points. A typical illustration of such a problem is given in Figure 1.

If we would like to calculate the correlation between two return series during, say, one day, we have to make sure that we have time series where the observations can be considered to origin from the same points in time. Take for example stock A and stock B and consider every second of the trading day.

Figure 1: Illustration of the problem



If these two stocks are not traded very frequently, we will not have *any* observations which origin from exactly the same time. Does it still make sense to calculate a correlation between the returns of them? Of course it does! It is just that we have to assume a relationship between different points in time in order to have quantities which we consider realistic to assume coming from the same points in time. An eloquent way of dealing with the situation with independence between the processes governing the observations and the time stamps is presented in De Jong and Nijman (1997). An important issue that will be emphasized in the present paper, is the possibility that the processes governing the observations and the time stamps are dependent. An example of this is when one consider a stock price. Imagining an underlying price, only observed occasionally. A possible reason for dependency is that an observation is more likely to occur if this underlying price make a large movement than if it makes a small one. This would imply that the duration between two trades would, on average, be smaller if the underlying price make large movements than if it makes small ones. This, in turn, implies a dependency between the two processes. In the next section, assumptions on the data generating process are stated. In Section 3 the problem studied in the paper is formulated. Section 4 reviews three different ways of doing this from a rather heuristic point of view by means of three interpolation methods. In Section 5 the consequences for the estimates of covariances and correlations when the interpolation methods in Section 4 are used, are studied. A method to calculate correlations directly, using the assumption that the returns are martingale differences and allowing for a dependency between the processes governing the observations and time stamps, is presented in Section 6. It is furthermore compared with a previous method (De Jong and Nijman, 1997) that does not allow for the possibility of such a dependency. In Section 8, the properties of the estimators are studied for cases where it is not possible to calculate them analytically. The conclusions are summarized in Section 9.

2 Assumption on the data generating process (DGP)

The application mainly thought of in this paper is the relationship between two financial asset prices and therefore, the type of processes that I will consider are of the type

$$\begin{cases} x_t = \mu_x + x_{t-1} + \delta_t \\ y_t = \mu_y + y_{t-1} + \varepsilon_t \end{cases} \tag{1}$$

where the error term processes$\{\delta_t\}$ and $\{\varepsilon_t\}$ is a simultaneously covariance stationary process with covariance function

$$Cov(\delta_t, \varepsilon_{t-k}) = \begin{cases} \gamma_k \text{ if } -K \leq k \leq K \\ 0 \text{ otherwise} \end{cases},$$

possibly with conditional heteroskedasticity. $\{\delta_t\}$ can always be written

$$\delta_t = \sum_{k=0}^{\infty} g_k a_t \tag{2}$$

where

$$\sum_{k=0}^{\infty} g_k^2 < \infty \tag{3}$$

and $\{a_t\}$ in turn can be written

$$a_t = \eta_t \sigma_t \tag{4}$$

where

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i a_{t-i} + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{5}$$

and $\{\eta_t\}$ is a sequence of independent identically distributed stochastic variables. The statements about $\{\delta_t\}$ above is also true for $\{\varepsilon_t\}$. In this setting, a natural example of x and y is that they represent the logs of two asset prices.

Finally, assume that irregularly spaced observations on two such time series which are correlated $x_{s_1}, x_{s_2}, ..., x_{s_{T_x}}$ and $y_{t_1}, y_{t_1}, ..., y_{t_{T_y}}$ are available and let $T = max(T_x, T_y)$

3 Description of the problem

In order to give a sense of the problem at hand we consider the first differences of the observed data

$$\begin{cases} \Delta x_i = \mu_x d_i^x + \sum_{s=s_{i-1}+1}^{s_i} \delta_s \\ \Delta y_j = \mu_y d_j^y + \sum_{t=t_{j-1}+1}^{t_j} \varepsilon_t \end{cases}$$

where $\Delta x_i = x_{s_i} - x_{s_{i-1}}$ and $\Delta y_i = y_{t_j} - y_{t_{j-1}}$. We also define the durations between observations $d_i^x = s_i - s_{i-1}$ and $d_j^y = t_j - t_{j-1}$. The reasons for this notation is that we want to study the quantity that later will be used to estimate the covariances and correlations, namely $Cov(\Delta x_i, \Delta y_j)$. Furthermore, the first differences of the data will be referred to as returns using finance terminology. The covariance can now be written as

$$Cov(\Delta x_i, \Delta y_j) = \mu_x \mu_y Cov\left(d_i^x, d_j^y\right) + \mu_x Cov\left(d_i^x, \sum_{t=t_{j-1}+1}^{t_j} \varepsilon_t\right)$$
$$+\mu_y Cov\left(\sum_{s=s_{i-1}+1}^{s_i} \delta_s, d_j^y\right) + Cov\left(\sum_{s=s_{i-1}+1}^{s_i} \delta_s, \sum_{t=t_{j-1}+1}^{t_j} \varepsilon_t\right)$$

Here we can see that this quantity is determined by three sources, the covariance between the durations of the two series, the covariance between the returns of the two series and by the covariance between the duration of one series and the return of the other.

Furthermore, if we look at the expression of the variance of a return, $Var(\Delta x_i)$, we see that also this is affected by the duration.

$$Var(\Delta x_i) = \mu_x^2 Var(d_i^x) + Var\left(\sum_{s=s_{i-1}+1}^{s_i} \delta_s\right) + 2\mu_x Cov\left(d_i^x, \sum_{s=s_{i-1}+1}^{s_i} \delta_s\right)$$

The effect on estimation of the last term of this expression will be studied later in this paper.

4 Interpolation methods

The obvious quick-fix for irregularly spaced data is to interpolate between observations of the two series in order to obtain data which, seemingly, origin from the same points in time. This seems like a natural thing to do. Most financial asset prices, after all, behave in a rather smooth manner so given reasonably frequent observations, an interpolation appear harmless. In the following, three different interpolation methods will be investigated in terms of how they affect estimation of covariances and correlations.

4.1 Step function

Since a process of the type described above is a martingale process, i.e. it has the property

$$E[x_t|\mathcal{F}_{t-k}] = x_{t-k} \tag{6}$$

where $\mathcal{F}_t = \{x_0, x_1, ..., x_t\}$, the guess of the type "the process does not change between observations" seems sensible. From a forecasting point of view, this is

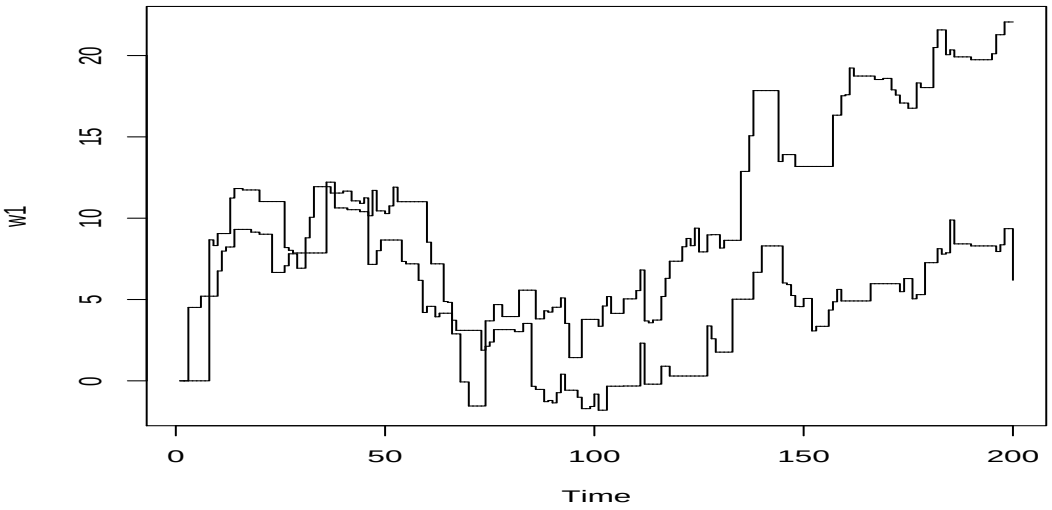


Figure 2: Step function interpolation

arguably the only reasonable guess since $\{a_t\}$ is assumed to be an unpredictable process. A typical plot of two such processes can be seen in Figure 2 where horizontal lines of length larger than one unit of time is the consequence of this type of interpolation.

4.2 Linear interpolation

If we are not interested in forecasting, we could use values located after the point in time we want to interpolate a value for as well as values before. A common approach here is to interpolate linearly in time. Say that we have a gap in our time series between the times t and $t+m$. A linear interpolation is performed so that the value x_{t+k}, where $k < m$ is estimated by

$$\widehat{x}_{t+k} = x_t + k\frac{x_{t+m} - x_t}{m} \tag{7}$$

A graph of two interpolated series is plotted in Figure 3. The same data as in Figure 2 is used.

4.3 Spline interpolation

The third interpolation method that will be considered is the interpolating cubic spline. This method produces more variation between the observations than the step function- and linear interpolation does. A cubic spline basis in s consists

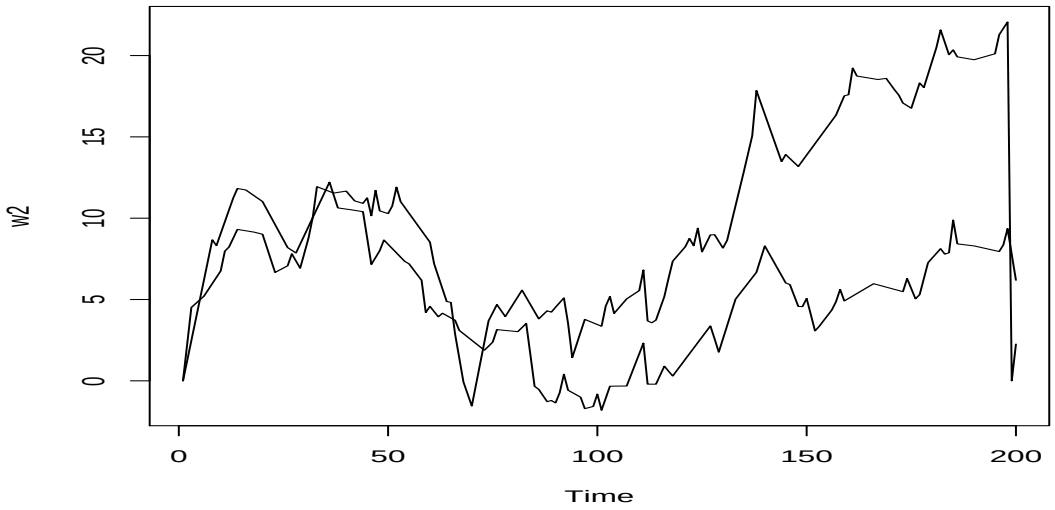


Figure 3: Linear interpolation

of the functions $1, s, s^2, s^3, (s-\xi_1)_+, ..., (s-\xi_b)_+$ where the ξ's are called *knots* and are all located between (or at) the minimum and maximum of the observed argument values, s_{min} and s_{max}. The function $(s)_+$ gives the maximum of s and zero. The data points x_{s_1},...,$x_{s_{T_x}}$ are then fitted to the basis with the ordinary least square method. The resulting trajectory of x seen as a function of s then obtain a continuous second order derivative. This property is contradictory with much of asset pricing theory which often use models based on Brownian motions which trajectories do not even have first order derivatives. However, for the purpose of estimating correlations and covariances, I will investigate this method as well as the other two interpolation methods in the sequel of this paper. The same data as for the two other interpolation techniques have been used in Figure 4 to exemplify the cubic spline interpolation.

5 Consequences under different interpolation methods

It is obvious, just by considering figures 2, 3 and 4, that a calculation of variance of or covariance between the first differences of two series generated by a process within the class (1)-(5), will depend on how the interpolation is made. In this section, I will consider these differences. The process under study are denoted

$$x_t = x_{t-1} + \delta_t \tag{8}$$

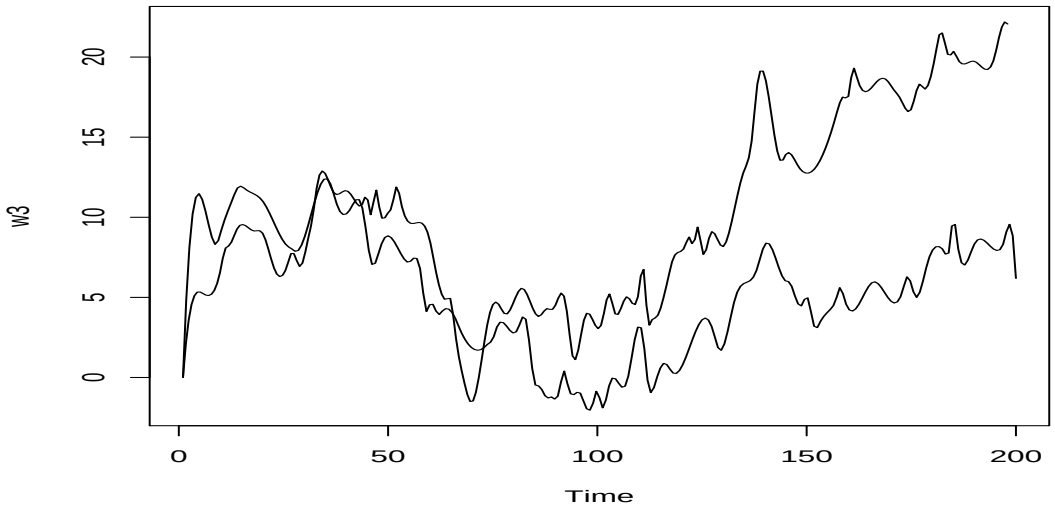


Figure 4: Cubic spline interpolation

and

$$y_t = y_{t-1} + \varepsilon_t \tag{9}$$

where the covariance and correlation between the white noise processes δ_t and ε_t are γ and ρ, respectively. The variances of δ_t and ε_t are σ_δ^2 and σ_ε^2. The problem is thus to estimate ρ. Assume, for simplicity, that the number of time points are the same for the two variables. Some stylized examples of irregularly spaced observations will be studied by imposing one "missing value". For this, we have observations say $x_1, ..., x_{k-1}, NA, x_{k+1}, ..., x_T$, where NA stands for missing value and $y_1, ..., y_{k-1}, y_k, y_{k+1}, ..., y_T$. The task is now to estimate ρ by

$$r = \frac{\sum_{t=1}^{T} \Delta x_t \Delta y_t}{\sqrt{\sum_{t=1}^{T} (\Delta x_t)^2 \sum_{t=1}^{T} (\Delta y_t)^2}} \tag{10}$$

where some values of Δx are substituted by interpolated values. Correlations between two observations at the exact same time point, is ρ and causes no particular problems. Therefore, correlation between an observed value of one variable and an interpolated value of the other will be considered. Since we are considering the returns and thereby the first differences of $\{x_t\}$ and $\{y_t\}$ there will for the missing value x_k be two missing values in $\{\Delta x_t\}$, Δx_k and Δx_{k+1}.

5.1 Step function interpolation

Since x_k is missing, the first differences that have to be replaced are $\Delta\hat{x}_k = 0$ and $\Delta\hat{x}_{k+1} = x_{k+1} - x_k$. These can, for the step function interpolation, be written $\Delta\hat{x}_k = 0$ and $\Delta\hat{x}_{k+1} = \delta_k + \delta_{k+1}$, respectively. Considering cross-products of these observations with the ones in the y-series first we observe that $Cov(\Delta\hat{x}_k, \Delta y_k) = Cov(\Delta\hat{x}_k, \Delta y_{k+1}) = 0$ and $Cov(\Delta\hat{x}_{k+1}, \Delta y_k) = Cov(\Delta\hat{x}_{k+1}, \Delta y_{k+1}) = \gamma$. Additionally, the variances of these observations are zero and $2\sigma_\delta^2$, respectively. Thus, the expected value of the estimator

$$\hat{\gamma}_{STEP} = \frac{1}{T-1}\sum_{t=2}^{T}\Delta x_t \Delta y_t$$

is

$$E(\hat{\gamma}_{STEP}) = \frac{T-2}{T-1}\gamma \tag{11}$$

Furthermore

$$Var(\Delta\hat{x}_k) = 0$$

and

$$Var(\Delta\hat{x}_{k+1}) = 2\sigma_\delta^2$$

implying

$$E(\hat{\rho}_{STEP}) \approx \frac{T-2}{T-1}\rho \tag{12}$$

5.2 Linear interpolation

According to the linear interpolation scheme we "estimate" x_k by

$$\hat{x}_k = \frac{x_{k-1} + x_{k+1}}{2}$$

which implies

$$\Delta\hat{x}_k = \Delta\hat{x}_{k+1} = \frac{1}{2}(x_{k+1} - x_{k-1})$$

or

$$\Delta\hat{x}_k = \Delta\hat{x}_{k+1} = \frac{1}{2}(\delta_k + \delta_{k+1})$$

Thus

$$Var(\Delta\hat{x}_k) = Var(\Delta\hat{x}_{k+1}) = \frac{1}{2}\sigma_\delta^2,$$

$$Cov(\Delta\hat{x}_k, \Delta y_k) = Cov(\frac{1}{2}(\delta_k + \delta_{k+1}), \varepsilon_k) = \frac{1}{2}\gamma$$

and, according to the same argument

$$Cov(\Delta\hat{x}_{k+1}, \Delta y_{k+1}) = \frac{1}{2}\gamma$$

Since $Var(\Delta y_k) = Var(\Delta y_{k+1}) = \sigma_\varepsilon^2$, then

$$Corr(\Delta\hat{x}_k, \Delta y_k) = Corr(\Delta\hat{x}_{k+1}, \Delta y_{k+1}) = \frac{1}{\sqrt{2}}\rho$$

Consequently, the expectation of the estimator

$$\hat{\gamma} = \frac{1}{T-1}\sum_{t=2}^{T}\Delta x_t \Delta x_t$$

is

$$E(\hat{\gamma}) = \frac{(T-2)}{T-1}\gamma$$

and the expectation of the variance estimator

$$\hat{\sigma}_\delta^2 = \frac{1}{T-1}\sum_{t=2}^{T}(\Delta x_t)^2$$

is

$$E(\hat{\sigma}_\delta^2) = \frac{(T-1)}{T}\sigma_\delta^2$$

implying

$$E(\hat{\rho}_{LIN}) \approx \sqrt{\frac{T-2}{T-1}}\rho. \tag{13}$$

5.3 Discussion on interpolation methods

In Section 8 the interpolation methods will be investigated in more general situations by means of a Monte Carlo study. Nevertheless, a brief discussion will here be made about the consequences of interpolation on covariance and correlation estimates. As formulas and show, the effect of a limited number of "missing values" will cause the estimates to be biased downwards. In addition, this will be more so for the step interpolation than for the linear interpolation. If there were m missing values instead of just and those were located with at least two observations between each one, the formulas corresponding to (12) and (13) would be

$$E(\hat{\rho}_{STEP}) \approx \frac{T-m-1}{T-1}\rho$$

and

$$E(\hat{\rho}_{LIN}) \approx \sqrt{\frac{T-m-1}{T-1}}\rho.$$

The situation with several missing values in a row is more difficult to analyze analytically and therefore this is done by means of Monte Carlo simulations in Section 8. A possible explanation for the relative superiority of the linear interpolation ,that will be investigated further in Section 8, is that it imposes a variation "between observations" that better corresponds to the DGP. The suspicion that even the linear interpolation does not capture the full variation of the DGP makes one believe that the cubic spline might do a better job.

6 Exact methods

Instead of interpolating the prices one can use the, reasonably weak, assumption that the return process is a martingale difference outlined in Section 2. By considering the observations $x_{s_1}, x_{s_2}, ...x_{s_{T_x}}$ and $y_{t_1}, y_{t_2}, ...x_{t_{T_y}}$ and their first differences, rewriting them as

$$\Delta x_i = \sum_{k=s_{i-1}-1}^{s_i} \delta_k$$

and

$$\Delta y_j = \sum_{l=t_{j-1}-1}^{t_j} \varepsilon_l$$

we can calculate $Cov(\Delta x_i, \Delta y_j)$ by evaluating which time intervals in the x- and the y-sequences that are overlapping each other.

$$Cov(\Delta x_i, \Delta y_j) = E(\Delta x_i \Delta y_j) = \sum_{k=-K}^{K} \gamma_k \chi_{ij}(k) \tag{14}$$

where

$$\chi_{ij}(k) = max(min(s_i, t_j + k) - max(s_{i-1}, t_{j-1} + k), 0)$$

where the zero occur when the intervals are not overlapping.

6.1 An unbiased method for $K = 0$

From (14), an unbiased estimator of γ can be derived regardless of dependence between the processes governing the observations and the time stamps or not.

$$\hat{\gamma} = \frac{1}{M} \sum_{i=2}^{T_x} \sum_{j=2}^{T_y} \frac{\Delta x_i \Delta y_j}{\chi_{ij}(0)} \tag{15}$$

just ignoring the terms where $\chi_{ij}(0) = 0$. M is the number of overlapping intervals. The variances for $\{x_t\}$ and for $\{y_t\}$ are estimated by

$$\hat{\sigma}_x^2 = \frac{1}{T_x - 1} \sum_{i=2}^{T_x} \frac{(\Delta x_i)^2}{\Delta s_i}$$

and

$$\hat{\sigma}_y^2 = \frac{1}{T_y - 1} \sum_{i=2}^{T_y} \frac{(\Delta y_i)^2}{\Delta t_i}.$$

Finally, the estimator of ρ is given by

$$\hat{\rho} = \frac{\hat{\gamma}}{\hat{\sigma}_x \hat{\sigma}_y} \tag{16}$$

Above, it was assumed that there are no correlation between δ_s and ε_t for $s \neq t$.

6.2 A method accounting for $K \neq 0$

(De Jong and Nijman, 1997; De Jong et al., 1998) presented an eloquent method which more directly uses equation (14) and thereby accounting for the possibility that $K \neq 0$, namely

$$Cov(\Delta x_i, \Delta y_j) = E(\Delta x_i \Delta y_j) = \sum_{k=-K}^{K} \gamma_k \chi_{ij}(k)$$

By creating $z_{ij} = \Delta x_i \Delta y_j$, the regression

$$z_{ij} = \sum_{k=-K}^{K} \gamma_k \chi_{ij}(k) + \xi_{ij}$$

where $E(\xi_{ij}) = 0$ can be used to estimate the γ_k's. The problem that occur here is that it implicitly assumes no dependence between the price and time stamp processes, manifested in no correlation between the $\chi(k)$'s and ξ_{ij}.

7 Estimating a simple regression with irregularly spaced data

An application of the analysis above could be to estimate a simple regression model for data observed irregularly and with different points in time for the x- and y- variables. This could e.g. be wished in order to obtain the β-value of a stock. In this case Δy_t would be the log-return of an individual stock and Δy_t the log-return of a market index. The model is

$$\Delta y_t = \alpha + \beta \Delta x_t + u_t, \tag{17}$$

where$\{u_t\}$ is white noise. The parameter of interest is

$$\beta = \frac{Cov(\Delta x_t, \Delta y_t)}{Var(\Delta x_t)}$$

and is estimated by plugging in the estimators of the nominator and denominator, but we could also estimate the intercept by using sample versions of the moments involved in

$$\alpha = E(\Delta y_t) - \beta E(\Delta x_t).$$

8 Simulation study

The simulation study performed in this section was made in order to understand a few things that would have been difficult, if not impossible, to investigate by analytic methods. Those are, how the estimators are affected by a dependence between the processes governing the observations and the time stamps, by a very large fraction of missing values realistic for an asset price if time is measured in seconds (this would be a very complicated combinatorial problem indeed) and by small samples. For illustrative purposes, the presentation here here refers to a stock market.

Regularly spaced data (intrinsic prices) from two random walks with different correlations between the error terms ($\rho = \pm 0.2, \pm 0.5, \pm 0.8$) was first generated. In order to create irregularly spaced series, two methods were exploited:

1. 20% of observations are kept (defined as trades) on the basis of two independent random drawings. This implies independence between the process driving returns and the process driving time of trade.

2. A change in the underlying, partly unobservable, process (intrinsic price) outside the quartile range has probability 0.6 of resulting in observation (trade). Otherwise, this probability is 0.2. This imposes a dependency between the two processes.

The number of replicates was 1000 and the number of observations of the underlying, partly unobserved, process studied were 100 and 1000, respectively, implying the actual number of observations are approximately 20 and 200, respectively. The simulation study was performed using the R-language (R Development Core Team, 2005). Both the interpolation methods and the exact methods were studied.

8.1 Estimation of the correlation coefficient

Table 1 show the Monte Carlo means for the case with independence. The columns show, from left to right, the number of observations of the underlying process, the true value of ρ and the Monte Carlo means of the five estimators. As can be seen, the results for the step function (no change) and linear interpolations methods are disastrous. Even for the sample size 1000, the estimators are severely biased towards zero. The results for the spline interpolation are, however, more promising. Some simulations with sample size 10000 indicated that the estimator might be consistent but this have to be studied more by analytical tools. Compared with the exact methods, however, this estimator comes out unfavourably. Concerning the two exact methods, there is no obvious systematic difference. The estimator 16 is by construction approximately unbiased so this should not be an issue. The results also indicate that also the OLS-based estimator is unbiased. The Monte Carlo standard deviations of the two exact estimators are similar, as can be seen in Table 2. The most interesting observation in this table is that the spline estimator have the smallest Monte Carlo standard deviations. This is difficult to explain and must be studied further.

T	ρ	No change	Linear	Spline	Exact Unbiased	Exact OLS
100	0.2	0.022	0.093	0.127	0.208	0.213
100	0.5	0.065	0.220	0.307	0.456	0.473
100	0.8	0.102	0.356	0.509	0.776	0.721
100	-0.2	-0.029	-0.086	-0.127	-0.196	-0.172
100	-0.5	-0.062	-0.222	-0.303	-0.468	-0.496
100	-0.8	-0.092	-0.362	-0.512	-0.812	-0.786
1000	0.2	0.022	0.094	0.172	0.207	0.205
1000	0.5	0.058	0.233	0.429	0.498	0.499
1000	0.8	0.090	0.373	0.690	0.802	0.788
1000	-0.2	-0.022	-0.091	-0.173	-0.200	-0.199
1000	-0.5	-0.059	-0.235	-0.432	-0.502	-0.496
1000	-0.8	-0.091	-0.373	-0.695	-0.798	-0.805

Table 1: Monte Carlo mean of correlation estimates for the case of independence and no dependence between value and time stamp process.

The table 3 and 4presents the Monte Carlo means and standard deviations when the processes governing the observations and time stamps are dependent in the sense explained above. The spline estimator still does a good job and have small standard deviations. The functionality of the OLS-based method now breaks down while the unbiased method, as expected, work properly. The standard deviations, again, are similar.

8.2 Estimation of parameters in a simple regression

A simulation exercise was also performed on a simple regression model, thought to resample the situation of estimating a β-value of a stock. In this case, only the exact estimators were used. The DGP here was equation 17 with parameters $\alpha = 1$and $\beta = (0.5, 1, 1.5)$. The results, presented in tables 5 and 6, show the same pattern as the results for the correlation coefficients. The results are similar for the two estimators when no dependence are present but when such dependence is part of the DGP, the results for the OLS-based method are unfavourable.

9 Conclusions

In this paper, two different strategies to estimate covariances and correlations between unsynchronised, irregularly spaced time series, given an underlying martingale process, have been investigated. Within those strategies, respectively three and two different modifications have been looked upon. The first strategy, interpolating between observations, turned out to be very dependent on the interpolation technique. Interpolation with a step function and linear

T	ρ	No change	Linear	Spline	Exact Unbiased	Exact OLS
100	0.2	0.107	0.153	0.351	0.515	0.509
100	0.5	0.116	0.145	0.327	0.519	0.465
100	0.8	0.122	0.136	0.304	0.520	0.456
100	-0.2	0.107	0.150	0.341	0.488	0.475
100	-0.5	0.116	0.146	0.337	0.511	0.500
100	-0.8	0.118	0.136	0.291	0.492	0.519
1000	0.2	0.032	0.050	0.183	0.165	0.162
1000	0.5	0.035	0.045	0.148	0.162	0.160
1000	0.8	0.039	0.044	0.104	0.166	0.185
1000	-0.2	0.031	0.047	0.182	0.171	0.167
1000	-0.5	0.036	0.048	0.157	0.167	0.171
1000	-0.8	0.039	0.044	0.111	0.166	0.185

Table 2: Monte Carlo standard deviation of correlation estimates for the case of independence between value and time stamp process.

T	ρ	No change	Linear	Spline	Exact Unbiased	Exact OLS
100	0.2	0.074	0.114	0.163	0.191	0.148
100	0.5	0.178	0.281	0.407	0.472	0.388
100	0.8	0.304	0.463	0.664	0.764	0.611
100	-0.2	-0.072	-0.114	-0.158	-0.192	-0.155
100	-0.5	-0.181	-0.286	-0.426	-0.483	-0.387
100	-0.8	-0.300	-0.464	-0.666	-0.780	-0.617
1000	0.2	0.073	0.114	0.185	0.192	0.149
1000	0.5	0.187	0.289	0.466	0.485	0.375
1000	0.8	0.303	0.469	0.756	0.785	0.600
1000	-0.2	-0.074	-0.113	-0.178	-0.191	-0.147
1000	-0.5	-0.186	-0.288	-0.473	-0.482	-0.376
1000	-0.8	-0.304	-0.469	-0.757	-0.784	-0.599

Table 3: Monte Carlo mean of correlation estimates for the case of dependence between value and time stamp process.

T	ρ	No change	Linear	Spline	Exact Unbiased	Exact OLS
100	0.2	0.105	0.113	0.350	0.243	0.238
100	0.5	0.105	0.104	0.307	0.240	0.263
100	0.8	0.114	0.103	0.231	0.236	0.269
100	-0.2	0.101	0.113	0.349	0.247	0.237
100	-0.5	0.106	0.106	0.295	0.235	0.238
100	-0.8	0.114	0.097	0.219	0.230	0.250
1000	0.2	0.031	0.036	0.182	0.080	0.083
1000	0.5	0.033	0.034	0.148	0.075	0.078
1000	0.8	0.037	0.031	0.084	0.074	0.079
1000	-0.2	0.032	0.034	0.181	0.074	0.079
1000	-0.5	0.033	0.033	0.144	0.074	0.080
1000	-0.8	0.036	0.032	0.083	0.075	0.079

Table 4: Monte Carlo standard deviation of correlation estimates for the case of dependence between value and time stamp process.

T	DEP	α	β	Unbiased α	OLS α	Unbiased β	OLS β
100	no	1	0.5	0.8235	0.9062	0.5260	0.5401
100	no	1	1.0	0.8300	0.9287	1.1861	1.1170
100	no	1	1.5	0.8398	0.9086	1.5513	1.5595
100	yes	1	0.5	0.9543	0.9702	0.4759	0.3725
100	yes	1	1.0	0.9582	0.9740	0.9929	0.7867
100	yes	1	1.5	0.9512	0.9647	1.5101	1.2355
1000	no	1	0.5	0.9863	0.9918	0.4778	0.4605
1000	no	1	1.0	0.9808	0.9882	0.9882	0.9722
1000	no	1	1.5	0.9803	0.9885	1.4804	1.5027
1000	yes	1	0.5	0.9975	0.9989	0.4865	0.3807
1000	yes	1	1.0	0.9968	0.9982	0.9759	0.7476
1000	yes	1	1.5	0.9970	0.9981	1.4648	1.1394

Table 5: Monte Carlo mean of parameter estimates in simple regression model for the case of independence between value and time stamp process.

T	DEP	α	β	Unbiased α	OLS α	Unbiased β	OLS β
100	no	1	0.5	0.4261	0.3896	2.4958	2.5506
100	no	1	1.0	0.4644	0.4342	2.4569	2.4985
100	no	1	1.5	0.4721	0.4624	2.4103	2.5197
100	yes	1	0.5	0.1795	0.1827	0.7783	0.9172
100	yes	1	1.0	0.1996	0.2008	0.8042	0.9180
100	yes	1	1.5	0.2423	0.2321	0.8796	1.0374
1000	no	1	0.5	0.0614	0.0642	0.7036	0.8076
1000	no	1	1.0	0.0731	0.0741	0.7332	0.8290
1000	no	1	1.5	0.0874	0.0888	0.7854	0.8902
1000	yes	1	0.5	0.0463	0.0460	0.2361	0.2882
1000	yes	1	1.0	0.0533	0.0526	0.2428	0.2941
1000	yes	1	1.5	0.0666	0.0643	0.2538	0.3119

Table 6: Monte Carlo standard deviation of parameter estimates in simple regression model for the case of dependence between value and time stamp process.

interpolation gave catastrophic results while a cubic spline worked better in the Monte Carlo study performed. The hypothesis of this author is that the reason is the inability of interpolation methods to capture the true variation of the underlying martingale process. In this respect, the cubic spline does the best job of the investigated methods. Performing much better, while more computationally expensive, are the two methods based directly on the martingale assumption. This might not be a surprise since they explicitly exploit the underlying DGP. The main contribution of the paper is the method which is unbiased even when there is a dependency between the processes governing the observations and the time stamps, a situation relevant for applications to financial markets.

References

Campbell, J., A. Lo, and A. MacKinlay (1997): *The econometrics of financial markets*, Princeton University Press.

Dacorogna, M., R. Gencay, U. Mucller, R. Olsen, and O. Pictet (2001): *An Introduction to High-Frequency Finance*, Academic Press.

De Jong, F., R. Mahieu, and P. Schotman (1998): "Price discovery in the foreign exchange market: an empirical analysis of the yen/dmark rate," *Journal of International Money and Finance*, 17, 5–27.

De Jong, F. and T. Nijman (1997): "High frequency analysis of lead-lag relationships between financial markets," *Journal of Empirical Finance*, 4, 259–277.

R Development Core Team (2005): *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.