Estimating The Covariance Matrix From Unsynchronized High Frequency Financial Data

Bin Zhou
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142

#3807

April, 1995

Estimating The Covariance Matrix From Unsynchronized High Frequency Financial Data

Bin Zhou

Sloan School of Management

Massachusetts Institute of Technology

Cambridge, MA 02139

Abstract

This paper proposes an estimator of the covariance matrix of currencies using unsychronized and noisy high frequency observations. The estimator allows us to estimate the covariance matrix over a shorter time interval with more accuracy. The estimator is not f-consistent when there are so-called observation noises. Increasing observation frequency infinitely does not always increase the accuracy of the estimation. Optimal observation frequency is dependent on the ratio of the total volatility over the noise level. Daily covariance matrices of three exchange rates are calculated to demonstrate the methodology. The empirical results show that the correlations of the three currencies are strong but various over time.

Key Words: f-consistency; observation noise; volatility.

1 Introduction

The estimation of the covariance matrix of financial prices is necessary in portfolio optimization and risk management. Besides sample covariance, many other estimators have been proposed (Stein 1975, Dey and Srinivasan 1985). However, estimating the covariance matrix from daily data can have serious problems. Jobson and Korkie (1980) indicated that, in some cases, it is better to use the identical matrix instead of the sample covariance matrix in the portfolio selection. The problem is that the number of observations is not enough to estimate all entries of a big covariance matrix. To get around the problem, one may want to collect more data over longer time interval. However, the changing condition of markets may prevent us to do so. Another approach is to impose constrains on the covariance matrix to reduce the number of free parameters (Frost and Savaino, 1986). The constrain may be subjective and not reflect the reality of the market. This paper explores another possibility of using high frequency data. Because of fast-growing computer power, data is now available in ultra-high frequency, such as tick-by-tick. Exchange rates, for example, can easily have over one million observations in one year.

There are several difficulties in using high frequency data. The first one is the issue of unsynchronized time. The price or quote of each currency or stock comes at different times. The second difficulty is so-called observation noise

(Zhou 1995a). The high frequency time series can be viewed as observations from a continuous process with observation noises:

$$S(t_i) = P(t_i) + \epsilon_{t_i}, \quad t_i \in [a, b], \tag{1}$$

where $P(t)$ is a diffusion process

$$dP(t) = \mu(t) + \sigma(t)dW_t. \tag{2}$$

The $P(t)$ is referred to as a signal process. The noise ϵ_{t_i} only occurs at the time of observation. ϵ_{t_i} behaves irregularly and no distributional assumptions have been made about the noises. The representation captured many characteristics, such as strong negative autocorrelations in high frequency observations. The present of observation noises can cause great difficulties in constructing f-consistent (Zhou 1995b) estimators. For an estimator without f-consistency, increasing observation frequency infinitely can do more harm than good.

In the next section, I will concentrate on constructing the estimators for the covariance matrix of unsynchronized financial time series. Without loss of generality, I will only discuss estimating the covariance of two financial time series. The variance estimators can be found in Zhou (1995a,1995b). In the last section, I will gives the estimates of daily covariance and daily correlation matrix of three exchange rates.

2 Estimating The Covariance Using Unsynchronized Data

Suppose that two time series $\{S_X(t_i)\}$ and $\{S_Y(s_j)\}$ are observations from two processes of (1) with observation noise. To further simplify the problem, I assume that the diffusion process $P(t)$ is a Brownian motion with a time deformation:

$$\begin{cases} S_X(t_i) = \mu_X(t_i) + W_X(\tau_X(t_i)) + \epsilon_X(t_i), & a \le t_i \le b \\ S_Y(s_j) = \mu_Y(s_i) + W_X(\tau_Y(s_i)) + \epsilon_Y(s_i), & a \le s_i \le b. \end{cases}$$

$\epsilon_X(t_i)$, $\epsilon_Y(s_i)$ $W_X(\tau)$ and $W_Y(\tau)$ are all independent. The parameter of interest is the covariance of two Brownian motions $W_X(\tau)$ and $W_Y(\tau)$ over time interval $[a, b]$:

$$c(a, b) = \mathrm{Cov}(W_X(b) - W_X(a), W_Y(b) - W_Y(a)) \tag{3}$$

Two processes $W_X(\tau)$ and $W_X(\tau)$ are assumed to have no leading effect, i.e., for $\tau_1 \le \tau_2 \le \tau_3 \le \tau_4$

$$\begin{cases} \mathrm{Cov}(W_X(\tau_1) - W_X(\tau_2), W_Y(\tau_3) - W_Y(\tau_4)) = 0, & \text{and} \\ \mathrm{Cov}(W_Y(\tau_1) - W_Y(\tau_2), W_X(\tau_3) - W_X(\tau_4)) = 0. \end{cases} \tag{4}$$

Under assumption (4),

$$c(s, t) + c(t, u) = c(s, u), \ s < t < u.$$

Without making any assumption about the regularity of time sequence $\{t_i\}$ and $\{s_j\}$ and the regularity of the covariance of each pair of $S_X(t_i) - S_X(t_{i-1})$ and $S_Y(t_j) - S_Y(t_{j-1})$, it is very difficult to construct a maximum likelihood estimator of the covariance over time $[a, b]$ using observations within the interval. Instead, I propose the following unbised estimator:

$$\hat{c}(a,b) = \sum_{a \le s_j \le b} (S_X(t_{i^+(j)}) - S_X(t_{i^-(j-1)}))(S_Y(s_j) - S_X(s_{j-1})), \qquad (5)$$

where

$$i^+(j) = \min\{i : t_i \ge s_j\} \text{ and } i^-(j) = \max\{i : t_i \le s_j\}.$$

Notice that the subscripts i and j are interchangeable. That is, estimator (5) can also be written as

$$\hat{c}(a,b) = \sum_{a \le t_i \le b} (S_X(t_i) - S_X(t_{i-1}))(S_Y(s_{j^+(i)}) - S_X(s_{j^-(i-1)})), \qquad (6)$$

When we have synchronized times, the estimator is simply the sample covariance. To save computation time, series $\{S_Y(s_j)\}$ should have fewer data points.

Theorem 1 *The estimator (5) is unbiased, i.e.,*

$$\mathbf{E}\hat{c}(a,b) = c(a,b)$$

The proof is straight forward and can be found in the Appendix.

There are no distributional assumptions being made about the noises. They can be exotic or autocorrelated. The time deformation function $\tau_X(t)$ and $\tau_Y(s)$ do not need to be known. Therefore, the volatilities in each tick do not have to be constant or given. Without the presence of noises, the estimator is f-consistent. However, when there are nonnegligible noises, the variance of the estimator diverges.

Theorem 2 *Let* $Z_X(t_{i+(j)}) = W_X(t_{i+(j)}) - W_X(t_{i-(j-1)})$ *and* $Z_Y(s_j) = W_Y(s_j) - W_Y(t_{j-1})$. *If all* $\epsilon_X(t_i)$ *and* $\epsilon_Y(s_j)$ *are zeros and* $\max_i varZ_X(t_{i+(j)}) \to 0$, *then*

$$var(\hat{c}(a,b)) = \sum_{a \le s_j \le b} var(Z_X(t_{i+(j)}) Z_Y(s_j)) \to 0. \tag{7}$$

Otherwise,

$$var(\hat{c}(a,b)) \ge \sum_j var(\epsilon_X(t_{i+(j)})) var(\epsilon_Y(s_j)). \tag{8}$$

The proof is given in the Appendix.

If the majority of noises is nonzero, the right-hand side of equation (8) approaches infinity as the observation frequency increases. On the other hand, if the frequency is too low, the right-hand side of equation (7) stays high. There is an optimal observation frequency to minimize the variance of the estimator (5). To find such an optimal observation frequency without assuming any regularity of time sequences $\{t_i\}$ and $\{s_j\}$ is complicated. To get some ideas

about this optimal observation frequency, I discuss only the following simple example:

1. Time series $S_X(t_i)$ have size $m+1$ and $S_Y(s_j)$ have size $n+1$, $m = kn$.

$$s_{j-1} < t_{(j-1)k} < t_{(j-1)k+1}, ..., < t_{jk-1} < s_j, i = 1, .., n$$

except

$$t_0 = a, t_m = b, s_0 = a, s_n = b.$$

It is easy to see that $i^+(j) = jk$ and $i^-(j-1) = (j-1)k - 1$.

2. Let $Z_X(t_i) = W_X(t_i) - W_X(t_{i-1})$ and $Z_Y(s_j) = W_Y(s_i) - W_Y(s_{i-1})$. The variances of signal changes are all constant

$$var(Z_X(t_i)) = \frac{\sigma_X^2}{m} \text{ and } var(Z_Y(s_j)) = \frac{\sigma_Y^2}{n}$$

where $\sigma_X^2 = \tau_X(b) - \tau_X(a)$ and $\sigma_Y^2 = \tau_Y(b) - \tau_Y(a)$.

3. The noises have no autocorrelations and

$$\text{var}(\eta_X(t_i)) = \eta_X^2 \text{ and } \text{var}(\eta_Y(s_i)) = \eta_Y^2.$$

4.

$$\text{var}(Z_X(t_i)Z_Y(s_j)) = \alpha\sigma_X^2(t_i)\sigma_Y^2(s_j) = \alpha\frac{\sigma_X^2\sigma_Y^2}{n^2},$$

where α is a constant between 1 and 2;

Under these conditions, the variance of the estimator (5) is:

$$\text{var}(\hat{c}(a,b)) = \alpha(1+\frac{1}{k})\frac{\sigma_X^2\sigma_Y^2}{n} + 2\sigma_X^2\eta_Y^2 + 2\sigma_Y^2\eta_X^2 + 2\frac{\sigma_X^2\eta_Y^2}{m} + 4n\eta_Y^2\eta_X^2 \quad (9)$$

where α is a constant between 1 and 2. The proof of the equation is given in the Appendix.

From (9), the optimal observation frequency n is near $\sqrt{(\alpha(1+1/k)r_Xr_Y + 2r_X/k)/4}$, where $r_X = \sigma_X^2/\eta_X^2$ and $r_Y = \sigma_Y^2/\eta_Y^2$, the signal-to-noise ratios. When the size of one series is much bigger than the size of the other one, the optimal observation frequency is near $\sqrt{\alpha r_X r_Y/4}$. The optimal observation frequency is proportional to the signal-to-noise ratio. When there is high level of noises, high frequency data, such as tick-by-tick, may have too many data points to use the estimator (5). Of course we can throw out some data points. However, to utilize all the data points, I propose the following estimator. Again assume that $\{S_X(t_i)\}$ has m data points and $\{S_Y(s_i)\}$ has n data points, $n < m$. n is k times larger than the optimal observation frequency. I can first use $S_Y(t_j), j = 0, k, 2k, ...$ to estimate the covariance, then use $S_Y(t_j), j = 1, k+1, 2k+1, ...$ to estimate the covariance. Finally, I average these k estimates. In summary,

$$\hat{c}_k(a,b) = \frac{1}{k}\sum_{p=1}^{k}\sum_{j=p(k)n}(S_X(t_i^+(j)) - S_X(t_{i^-(j-k)}))(S_Y(s_j) - S_X(s_{j-k})), \quad (10)$$

This estimator is still not f-consistent. However, the upper bound of the

variance of the estimator (10) becomes finite when the observation frequency approaches to infinity.

3 Estimating Covariance and Correlation Matrices of Three Exchange Rates

In this section, I will estimate the daily covariance and correlation matrices of exchange rates from tick-by-tick data. The high frequency data is provided by Olsen & Associates. It is referred to as the HFDF93 data set. It contains spot rate quotes of the Deutsche mark and US dollar (DEM/USD), the Japanese yen and US dollar (JPY/USD) and the Deutsche mark and Japanese yen (JPY/DEM) from October 1, 1992 to September 30, 1993. The data is recorded from the Reuters screen. DEM/USD are the most active currencies traded in the market followed by JPY/USD. Cross-rates, like JPY/DEM, are much less active. In this paper only the bid prices are used because the bid price is quoted in its entirety. The ask price is quoted by the last two or three digits only. Interpreting ask price by computer is often troublesome. Since the data is a non-binding quote, the level of observation noise is very high. The basic statistics of the returns of the three exchange rates are listed in Table 1.

Table 1: Summary Statistics of Tick-by-tick Returns

Series	n	Min.	Max.	Mean	sd	Skew.	Kurt.	ρ
DEM/USD	1472032	−.0066	.00544	9.92e-8	2.67e-4	.017	6.87	−.451
JPY/USD	570689	−.0105	0.0104	−2.18e-7	3.72e-4	−.029	32.17	−.425
JPY/DEM	158958	−.0098	0.0100	−1.70e-6	3.50e-4	−.385	23.26	−.107

ρ: the first order autocorrelation.

To examine if there are any lagged correlations among the three exchange rates, cross-correlations of daily returns are calculated. Figure 1 shows that three exchange rates have significant correlations, but there are no lagged correlations.

To calculate the correlation matrix, the following volatility estimator is used (Zhou 1995a):

$$\hat{\sigma^2}_k = \frac{1}{k}\sum_{p=1}^{k}\sum_{j=p(k)n}(S(t_j) - S(t_{j-k}))(S(t_{j+k}) - S(t_{j-2k})). \tag{11}$$

and the level of noise is estimated by the sample autocovariance

$$\eta^2 = -\frac{1}{kn}\sum_{p=1}^{k}\sum_{i=p(k)n}(S(t_i) - S(t_{i-k}))(S(t_{i-k}) - S(t_{i-2k})). \tag{12}$$

The estimates of the annual volatilities, the variances of the signals, and the noise level η^2 are listed in Table 2.

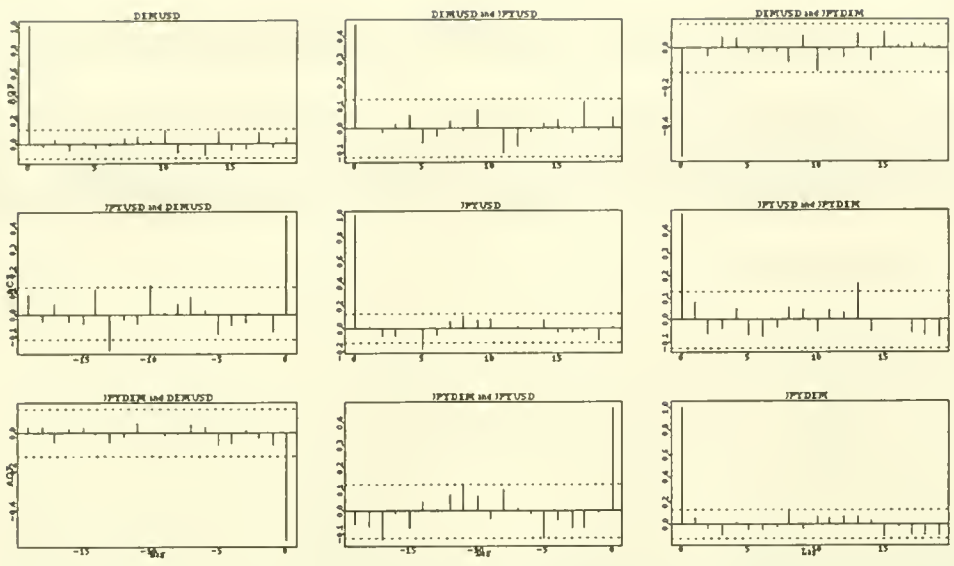

Figure 1: Cross Correlation of Three Exchange Rates Using Daily Returns.

Table 2: Estimation of Signal and Noise Levels

Series	n	k	Ann. vol.	η^2	S-N r
DEM/USD	1472032	5	.0148	2.86e-8	517482
JPY/USD	570689	3	.0153	5.35e-8	285981
JPY/DEM	158958	1	.0153	1.30e-8	1176923

From Table 2, I can get a rough estimate of the optimal observation frequency for each pair of exchange rates. For DEM/USD and JPY/USD, the rough estimate is about 260,000. Compared to the size of the smaller series JPY/USD, it is less than one half. I choose $k = 3$ in formula (10) to estimate the covariance. Using a similar argument, $k = 1$ is chosen in the two other pairs of exchange rates. The daily covariances of the three pairs of exchange rates are given in Figure 2.

Since the market volatility changes over time, the change of the covariance is affected by changing volatility; therefore a correlation matrix may be more desirable in some cases. Using volatility estimator (11), daily volatilities are calculated for each exchange rate. Daily correlations of the three exchange rates are plotted in Figure 3.

There are several interesting observations from Figure 3. First, the correlation between DEM/USD and JPY/USD are almost always positive. This indicates that the US dollar is a leading currency. When it moves, it moves in the same direction against both the Deutsche mark and the Japanese yen. The Deutsche mark has the same feature. During the first four and a half months of the time interval, both the US dollar and the Deutsche mark dominated the market. However, after February of 1993, the Japanese yen became a dominating currency. The role of each currency changes over a long time interval.

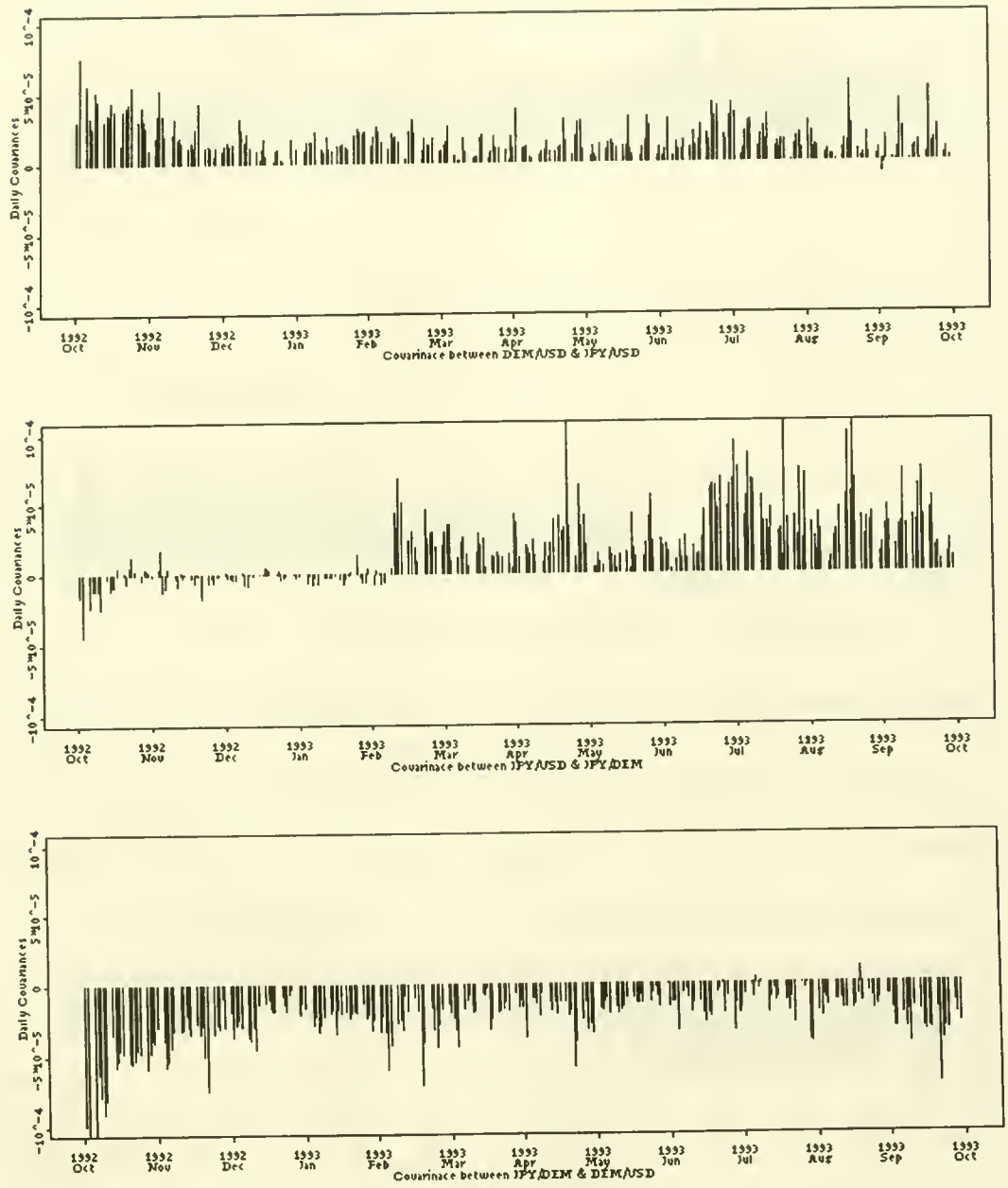

Figure 2: Daily Estimates of the Covariance of Three Exchange Rates.

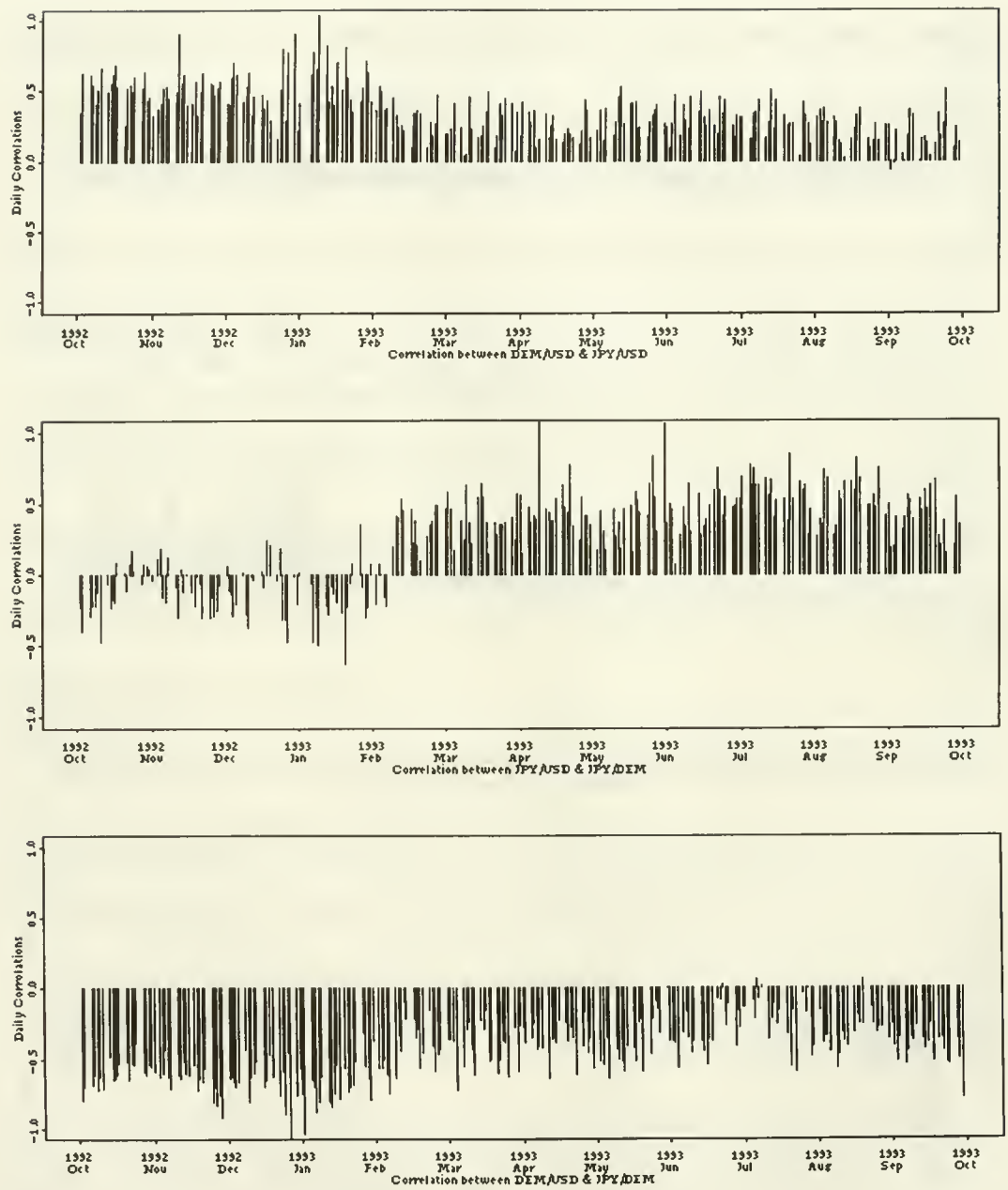

Figure 3: Daily Estimates of the Correlation of Three Exchange Rates.

However, they are relatively stable over short periods such as months.

Since the three exchange rates are dependent, the actual covariance matrix is singular. Buying one unit of DEM/USD and JPY/DEM and selling one unit of JPY/DEM end a neutral position. The empirical results confirmed that the minimum eigenvalue of estimated monthly covariance matrices are very close to zero for all twelve months.

4 Discussion

High frequency data provides us with enough data not only to estimate a large covariance matrix, but also to estimate the variance matrix over a short time interval. This is extremely beneficial in a fast changing market. It enables people to see the market change quickly and adjust their portfolio in time. Of course, using high frequency data is not without its difficulties. The nonsynchronized time causes great difficulty in getting a maximum likelihood estimator and the observation noise causes great difficulty in achieving an f-consistency. The covariance estimator proposed here is simple with few assumptions. However, if one is willing to impose more conditions, the estimator can be improved.

In the currency market, it is reasonable to assume no leading correlation (4) among exchange rates. This may not be true in the stock market when

small stocks are considered. Lo (1990) showed that there is asymmtretry in the stock market, i.e., big stocks may lead small stocks. The proposed estimator does not work in such a case. However, small stocks are often thinly traded; they are not the focus of this research.

Appendix:

i) Proof of Theorem 1, the unbiasness of estimator (5):

First, I write the change of prices as following:

$$\begin{cases} S_X(t_{i^+(j)}) - S_X(t_{i^-(j-1)}) = Z_X(t_{i^+(j)}) + \epsilon_X(t_{i^+(j)}) - \epsilon_X(t_{i^-(j-1)}), \text{ and} \\ S_Y(s_j) - S_Y(t_{j-1}) = Z_Y(s_j) + \epsilon_Y(s_j) - \epsilon_Y(s_{j-1}). \end{cases}$$

Then

$$\begin{aligned} \mathbf{E}\hat{c}(a,b) &= \mathbf{E}\sum_{a\le s_j\le b}[Z_X(t_{i^+(j)})Z_Y(s_j) + Z_X(t_{i^+(j)})\epsilon_Y(s_j) - Z_X(t_{i^+(j)})\epsilon_Y(s_{j-1}) \\ &\quad +\epsilon_X(t_{i^+(j)})Z_Y(s_j) + \epsilon_X(t_{i^+(j)})\epsilon_Y(s_j) - \epsilon_X(t_{i^+(j)})\epsilon_Y(s_{j-1}) \\ &\quad -\epsilon_X(t_{i^-(j-1)})Z_Y(s_j) - \epsilon_X(t_{i^-(j-1)})\epsilon_Y(s_j) + \epsilon_X(t_{i^-(j-1)})\epsilon_Y(s_{j-1})] \\ &= \sum_{a\le s_j\le b}\mathbf{E}Z_X(t_{i^+(j)})Z_Y(s_j) \\ &= \sum_{a\le s_j\le b}c(s_{j-1},s_j) \\ &= c(b,a) \end{aligned}$$

i) Proof of Theorem 2, the f-consistency of estimator (5):

Use the same notation as in Theorem 1 and let $\sigma_X^2(t_{i^+(j)})$=var$(Z_X(t_{i^+(j)}))$ and $\sigma_Y^2(s_i)$=var$(Z_Y(s_i))$. When there are no noises

$$\begin{aligned} \text{var}(\hat{c}(a,b)) &= \text{var}(\sum_{a\le s_j\le b}Z_X(t_{i^+(j)})Z_Y(s_j)) \\ &\le \sum_{a\le s_j\le b}\mathbf{E}(Z_X^2(t_{i^+(j)})Z_Y^2(s_j)) \\ &\le \sum_{a\le s_j\le b}\sqrt{\mathbf{E}Z_X^4(t_{i^+(j)})\mathbf{E}Z_Y^4(s_j)} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{a \le s_j \le b} \sqrt{3\sigma_X^4(t_{i+(j)}) 3\sigma_Y^4(s_j)} \\
&\le 3 \max_i \sigma_X^2(t_{i+(j)}) \sum_j \sigma_Y^2(s_j) \to 0.
\end{aligned}
$$

On the other hand, when there are noises,

$$
\mathrm{var}(\hat{c}(a,b)) \ge \mathrm{var}(\sum_{a \le s_j \le b} \epsilon_X(t_{i+(j)}) \epsilon_Y(s_j)).
$$

i) Proof of equation (9):

$$
\begin{aligned}
\mathrm{var}(\hat{c}(a,b)) &= \mathrm{var}(\sum_{i=1}^{n} [(Z_X(t_{(j-1)k}) + ... + Z_X(t_{jk})) Z_Y(s_j) \\
&\qquad + (Z_X(t_{(j-1)k}) + ... + Z_X(t_{jk})) \epsilon_Y(s_j) \\
&\qquad - (Z_X(t_{(j-1)k}) + ... + Z_X(t_{jk})) \epsilon_Y(s_{j-1}) \\
&\quad + \epsilon_X(t_{jk}) Z_Y(s_j) + \epsilon_X(t_{jk}) \epsilon_Y(s_j) - \epsilon_X(t_{jk}) \epsilon_Y(s_{j-1}) \\
&\quad - \epsilon_X(t_{(j-1)k-1}) Z_Y(s_j) - \epsilon_X(t_{(j-1)k-1}) \epsilon_Y(s_j) + \epsilon_X(t_{(j-1)k-1}) \epsilon_Y(s_{j-1})] \\
&= \alpha(1 + \frac{1}{k}) \frac{\sigma_X^2 \sigma_Y^2}{n} + 2\sigma_X^2 \eta_Y^2 + 2\frac{\sigma_X^2 \eta_Y^2}{m} + 2\sigma_Y^2 \eta_X^2 + 4n\eta_Y^2 \eta_X^2
\end{aligned}
$$

References

[1] Dey, D. K. and C. Srinivasan (1985), "estimation of a Covariance Matrix Under Stein's Loss." *annuals of Statistics,* **13**, 1581-1591

[2] Frost, P.A. and J.E. Savarino (1986), "An Empirical Bay's Approach to Portfolio Selection." *J. of Financial and Quantitative Analysis,* **21**, 293-305

[3] Goodhart, C.A.E. and L. Figliuoli (1991), "Every minute counts in financial markets." *J. of Inter. Money and Finance,* **10**, 23-52.

[4] Jobson, J. D. and B. Korkie (1980), "Estimation for Markowitz efficient portfolios." *JASA,* **75**, 544-554.

[5] Ledoit, Olivier (1994), "Portfolio Selection: Improved Covariance Matrix Estimation." Ph.D. Dissertation, Sloan School of Management, MIT

[6] Lo, Andrew and A.C. MacKinlay (1990), "When Are Contrarian Profits Due to Stock Market Overeaction?." *The Review of Financial Studies,* **3**, 175-205.

[7] Markowitz (1952), "Portfolio Selection." *J. of Finance,* **7**, 77-91.

[8] Stein, C. (1975) "Estimation of a Covariance Matrix." Rietz Lecture, 39th Annul Meeting IMS, Atlanta, GA.

[9] Zhou, Bin (1995a), "High Frequency Data and Volatility In Foreign Exchange Rates." JBES, in press.

[10] Zhou, Bin (1995b), "Estimating The Variance Parameter From Noisy High Frequency Financial Data." MIT Sloan School Working Paper.

[11] Zhou, Bin (1994), "Forecasting Foreign Exchange Rates Subject to Devolatilization." *Artificial Intelligence in the Capital Markets*, Ed. by Freedman, Klein & Lederman, Probus, Chicago.