Journal of Data Science **10**(2012), 345-362

Dynamic Co-movement Detection of High Frequency Financial Data

Mei-Hui Guo[1], Ching-An Liu[1] and Shih-Feng Huang[2*]
[1] *National Sun Yat-sen University and*
[2] *National University of Kaohsiung*

Abstract: In this study, we propose a pattern matching procedure to seize similar price movements of two stocks. First, the algorithm of searching the longest common subsequence is introduced to sieve out the time periods in which the two stocks have the same integrated volatility levels and price rise/drop trends. Next we transform the price data in the found matching time periods to the Bollinger Percent b data. The low frequency power spectra of the transformed data are used to extract trends. Pearson's chi-square test is used to assess similarity of the price movement patterns in the matching periods. Simulation results show the proposed procedure can effectively detect the co-movement periods of two price sequences. Finally, we apply the proposed procedure to empirical high frequency transaction data of NYSE.

Key words: Bollinger Percent, high frequency transaction data, longest common subsequence, pattern matching, power spectrum.

1. Introduction

In security markets, the stock price movements are closely linked to the market information. For example, the news on subprime mortgage crisis triggered a global financial crisis through 2007 and 2008. Drops occurred in virtually every stock market in the world. After the Federal Reserve took several steps to address the crisis, the stock markets have been gradually stable. For intraday trading, the finance literature highlights that the arrival of information over intradaily frequencies has also a strong impact on both prices and volatility and affects the security market activities. Traders in securities markets are often characterized in two groups, that is, informed and liquidity traders. Informed traders carry private information. Securities prices become more informative when there are

[*] Corresponding author.

more informed traders in the market and liquidity traders prefer to trade in an informative market than otherwise. Reaction of the traders to the same information on stocks from the same sector results in similar price movements, yet their reaction time might be different. The study of arrival of informed traders or asymmetric information is an important subject for microstructure analysis of financial market. Thanks for modern computer technology, ultra high frequency financial data such as transaction-by-transaction data now has become available and provide a rich source in studying intraday market microstructure dynamics. In this paper, we use the high frequency transaction data to investigate the similar price movement patterns of two stocks around information arrivals.

Pattern matching is an important subject in future movement prediction, rule discovery and computer aided diagnosis. In the literature, the longest common subsequence (LCS) method is widely used in bioinformatics for biological sequence alignment. The LCS problem (Hirschberg, 1975, Agrawal, Faloutsos and Swami, 1993, Bergroth, Hakonen and Raita, 2000, and Dacorogna, Gençay, Müller, Olsen and Pectet, 2001) is to find the longest subsequence common to all sequences in a set of sequences (often just two). It is a classical problem in computer science and has applications in many fields. For example, biologists can decide similarity of two DNA sequences by the length of their LCS, the Unix program "diff" compare two different versions of the same file by finding a LCS of the lines in the two files.

In this study, we propose a four stage procedure to search similar patterns for intraday high frequency transaction data. First, we apply the LCS method to sieve out the time intervals in which the two stocks have the same integrated volatility levels as well as the price rise/drop trends. Next, we transform the price data sieved out from the first step to the Bollinger Percent b data, then use the power spectrum to filter out the low frequency components. The fourth step is to assess similarity of the price movement patterns in the matching periods by Pearson's chi-square test. There are several advantages of the proposed approach. For example, the LCS algorithm heightens efficiency of searching periods of similar price patterns, the power spectrum are easily obtained by software package and the Pearson's chi-square test provides a powerful and objective test.

The remainder of the paper is organized as follows. In Section 2, we introduce the stock price models. In Section 3, the LCS method is introduced. In Section 4, the proposed method is introduced and simulation and empirical studies are performed. Conclusion is given in Section 5. Tables and figures are in the Appendix.

2. Model Assumptions

In real market high frequency transactions arrive randomly. Equi-spaced sample can be obtained via certain synchronization scheme such as the previous-tick interpolation scheme (Dacorogna, Gençay, Müller, Olsen and Pectet, 2001). In this study we assume the stock prices are available at equi-distance times $t_i = i\Delta$, $i = 1, \cdots, n$, where Δ denotes the length of the sampling interval. Let S_i^A and S_i^B denote the stock prices of Companies A and B at time t_i, respectively. Assume the log return of stock A has the following conditional normal distribution

$$\log(S_i^A/S_{i-1}^A) \sim N((\mu_A - \frac{\sigma_A^2}{2})\Delta,\ \sigma_A^2\Delta), \tag{2.1}$$

where μ_A is the annualized average return and σ_A^2 is the annualized volatility. The New York Stock Exchange trades for 6.5 hours per day from 09:30 AM to 16:00 PM. To simulate the real market, we generate 5201 equispaced stock prices in 6.5 hours, with sampling length $\Delta = \frac{1}{250} \times \frac{1}{5200} = 7.69 \times 10^{-7}$ (year). Divide the 6.5 hours into 26 non-overlapping 15-minute time interval denoted by $b_1, b_2, \cdots, b_{26}$, with 200 returns in each interval. In the first 10 intervals $b_1, b_2, \cdots, b_{10}$, the information receiving time of A lead B by 15 minutes. In the next 6 intervals, $b_{11}, b_{12}, \cdots, b_{16}$, the information receiving time of the two companies are synchronous. In the last 10 intervals, $b_{17}, b_{18}, \cdots, b_{26}$, the information receiving time of A lags B by 15 minutes. That is, we consider the following postulated models for Companies A and B:

$$S_i^B = \begin{cases} \alpha S_{i-200}^A + \beta + \varepsilon_i, & i = 1, \cdots, 2001, \\ \alpha S_i^A + \beta + \varepsilon_i, & i = 2002, \cdots, 3201, \\ \alpha S_{i+200}^A + \beta + \varepsilon_i, & i = 3202, \cdots, 5201, \end{cases} \tag{2.2}$$

where α and β are constants. Since time-varying heteroscedastic features are frequently observed in a financial time series, herein we assume the noise term ε_i comes from the following GARCH(1,1) model,

$$\begin{cases} \varepsilon_i &= \sigma_i \eta_i,\ \eta_i \sim N(0,1), \\ \sigma_i^2 &= \alpha_0 + \alpha_1 \varepsilon_{i-1}^2 + \beta_1 \sigma_{i-1}^2, \end{cases} \tag{2.3}$$

where α_0, α_1 and β_1 are positive constants and $\alpha_1 + \beta_1 < 1$. Figure 1 is the time plots of the generated stock prices of the companies A and B. We are interested in detecting the dynamic co-movements of the stocks A and B.

Since the stock price process is generally non-stationary, unless for cointegrated processes, regression models might result in spurious regression. Thus it is not suitable to be applied regression analysis in this study. Moreover due the nonlinear relationship between A and B, the linear correlation coefficient is not useful in this case either. Define the integrated volatility of interval b_i as

$$v_i = \sum_{j=1}^{200} \tilde{r}_{i,j}^2, \quad i = 1, 2, \cdots, 26,$$

where $\{\tilde{r}_{i,j}\}_{j=1}^{200}$ are the log returns in the time interval b_i. The correlations between the returns and integrated volatilities of stocks A and B are 0.240 and 0.062, respectively, which do not suggest significant linear relationship between the companies A and B. In the following section, we introduce the LCS method to search the time intervals in which the two stocks have similar integrated volatility and price rise/drop trends.

3. Longest Common Subsequence

A string $\mathbf{u} = u_1u_2\cdots u_k$ is called a subsequence of a string $\mathbf{v} = v_1v_2\cdots v_n$ if there is a mapping $F : \{1, 2, \cdots , k\} \to \{1, 2, \cdots , n\}$, $k \leq n$, such that $F(i) = l$ only if $u_i = v_l$ and F is a monotone strictly increasing function, that is, if $F(i) = p$, $F(j) = q$ and $i < j$, then $p < q$. For example, "coin" is a subsequence of "correlation". In addition, a string $\mathbf{u}$ is called a common subsequence of two strings $\mathbf{v}$ and $\mathbf{w}$ if $\mathbf{u}$ is a subsequence of both $\mathbf{v}$ and $\mathbf{w}$. Formally, we define the common subsequence of strings $\mathbf{v} = v_1v_2\cdots v_n$, and $\mathbf{w} = w_1w_2\cdots w_m$ as a sequence of positions in $\mathbf{v}$,

$$1 \leq i_1 < i_2 < \cdots < i_k \leq n$$

and a sequence of positions in $\mathbf{w}$,

$$1 \leq j_1 < j_2 < \cdots < j_k \leq m$$

such that the symbols at the corresponding positions in $\mathbf{v}$ and $\mathbf{w}$ coincide:

$$v_{i_t} = w_{j_t}, \quad t = 1, 2, \cdots , k.$$

For example, "eat" is a common to both "correlation" and "relationship". Finally, we define string $\mathbf{u}$ to be a longest common subsequence of string $\mathbf{v}$ and $\mathbf{w}$ if $\mathbf{u}$ is a common subsequence of $\mathbf{v}$ and $\mathbf{w}$ of maximal length. For example, "relation" is the longest common subsequence of "correlation" and "relationship".

The longest common subsequence problem can be solved by dynamic programming, which gives a way of making the solution more efficient. To do this, we introduce the following recursive solution. Define $s_{i,j}$ to be the length of an LCS between $v_1\cdots v_i$, the i-prefix of $\mathbf{v}$ and $w_1\cdots w_j$, the j-prefix of $\mathbf{w}$. Clearly, $s_{i,0} = s_{0,j} = 0$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$. One can see that $s_{i,j}$ satisfies the following recurrence:

$$s_{i,j} = \max \begin{cases} s_{i-1,j} + 0, \\ s_{i,j-1} + 0, \\ s_{i-1,j-1} + 1, & \text{if } v_i = w_j. \end{cases}$$

The first term corresponds to the case when v_i is not present in the LCS of the i-prefix of $\mathbf{v}$ and j-prefix of $\mathbf{w}$; the second term corresponds to the case when w_j is not present in this LCS; and the third term corresponds to the case when both v_i and w_j are present in the LCS. Note that the matching positions of the LCS $\{(i_1, j_1), (i_2, j_2), \cdots, (i_k, j_k)\}$ may not be unique. For example, suppose that the string $\mathbf{v}$ is **abcdbb** and the string $\mathbf{w}$ is **cbacbaaba**, then both **bcbb** and **acbb** are the LCSs with corresponding matching positions {(2,2),(3,4),(5,5),(6,8)} and {(1,3),(3,4),(5,5),(6,8)}, respectively. Figure 2 is an illustration of the matching positions of the LCSs of strings $\mathbf{v}$ and $\mathbf{w}$.

In this study, the LCS method is used to find the similar market reaction periods of two stocks to the intradaily information. As mentioned in the previous section, the normal trading hours (9:30am-4:00pm) are divided into 26 nonoverlapping periods each of length 15 minutes. Let v_i denote the integrated volatility in the ith time period b_i, that is $v_i = \sum_{j \in I_i} \tilde{r}_{i,j}^2$, where $\tilde{r}_{i,j}$ is the observed jth log return in the time period b_i. The price movements within a time period are classified into the following 8 categories by their integrated volatility levels and the price trends in the period :

$$s_i = \sum_{k=1}^{4} k \cdot I_{\{Q_{k-1} < v_i \le Q_k\}} \cdot \operatorname{sgn}\Big(\sum_{j=1}^{200} \tilde{r}_{i,j}\Big), \tag{3.1}$$

where sgn(x) is the sign function of x, I_A is the indicator function of the set A, $Q_0 = 0$, Q_k, $k = 1, 2, 3$, are the quartiles of $\{v_i\}_{i=1}^{26}$ and $Q_4 = \infty$. Thus s_i takes values in $\{\pm 1, \pm 2, \pm 3, \pm 4\}$. The classification criterion (3.1) considers two factors together, the level of the integrated volatility and the price trend in the time period b_i. In particular, if $\sum_{j=1}^{200} \tilde{r}_{i,j} = 0$, then

$$s_i = \begin{cases} \sum_{k=1}^{4} k \cdot I_{\{Q_{k-1} < v_1 \le Q_k\}}, & i = 1, \\ \sum_{k=1}^{4} k \cdot I_{\{Q_{k-1} < v_i \le Q_k\}} \operatorname{sgn}(s_{i-1}), & i \ge 2. \end{cases}$$

Let $\{s_i^A\}_{i=1}^{26}$ and $\{s_j^B\}_{j=1}^{26}$ denote the categorized sequences of the stock prices of companies A and B, respectively. We apply the LCS method to find the matching time intervals of $\{s_i^A\}_{i=1}^{26}$ and $\{s_j^B\}_{j=1}^{26}$, with the same integrated volatility levels and price trends.

Under Model (2.2) the true LCS of $\{s_i^A\}_{i=1}^{26}$ and $\{s_j^B\}_{j=1}^{26}$ is

$$C_{AB} = C_{AB}^{(1)} \cup C_{AB}^{(2)} \cup C_{AB}^{(3)}, \tag{3.2}$$

where

$$\begin{aligned}C_{AB}^{(1)} &= \{(1,2),(2,3),(3,4),(4,5),(5,6),(6,7),(7,8),(8,9),(9,10)\},\\ C_{AB}^{(2)} &= \{(11,11),(12,12),(13,13),(14,14),(15,15),(16,16)\},\\ C_{AB}^{(3)} &= \{(18,17),(19,18),(20,19),(21,20),(22,21),(23,22),(24,23),(25,24),\\ &\qquad (26,25)\}.\end{aligned}$$

The lengths of $C_{AB}^{(1)}$, $C_{AB}^{(2)}$, $C_{AB}^{(3)}$, and C_{AB} are 9, 6, 9 and 24, respectively.

In the following, we perform simulation study to investigate the LCS method for pattern matching of Model (2.2). Let $\hat{C}_{AB} = \{(i_h, j_h) : h = 1, 2, \cdots, k, 1 \leq i_1 < i_2 < \cdots < i_k \leq 26 \text{ and } 1 \leq j_1 < j_2 < \cdots < j_k \leq 26\}$ denote the LCS of the simulated sequences. Then $\hat{C}_{AB} = \mathit{CMP} \cup \mathit{ICMP}$ where $\mathit{CMP} \equiv \hat{C}_{AB} \cap C_{AB}$ and $\mathit{ICMP} \equiv \hat{C}_{AB} \backslash C_{AB}$, which represent the correct matching positions and incorrect matching positions, respectively. For example, if $\hat{C}_{AB} = \{$(1,2), (5,5), (6,7), (12,12), (13,14), (20,19), (22,22), (25,24)$\}$, compared with C_{AB} defined in (3.2), then we have $\mathit{CMP} = \{$(1,2), (6,7), (12,12), (20,19), (25,24)$\}$ and $\mathit{ICMP} = \{(5,5),(13,14),(22,22)\}$. Furthermore, let $\hat{C}_{AB}^{(1)} = \{(i,j) : (i,j) \in \hat{C}_{AB} \text{ and } i < j\}$, $\hat{C}_{AB}^{(2)} = \{(i,j) : (i,j) \in \hat{C}_{AB} \text{ and } i = j\}$, and $\hat{C}_{AB}^{(3)} = \{(i,j) : (i,j) \in \hat{C}_{AB} \text{ and } i > j\}$. Let $l(x)$ denote the length of a sequence x, then

$$\frac{l(C_{AB} \cap \hat{C}_{AB})}{l(\hat{C}_{AB})} = \frac{l(\mathit{CMP})}{l(\hat{C}_{AB})}$$

denotes the correct rate of the LCS method for seizing the matching positions. Similarly, the correct rate of the LCM method when A leads B is

$$\frac{l(C_{AB}^{(1)} \cap \hat{C}_{AB}^{(1)})}{l(\hat{C}_{AB}^{(1)})},$$

the correct rate when A and B are contemporaneous is

$$\frac{l(C_{AB}^{(2)} \cap \hat{C}_{AB}^{(2)})}{l(\hat{C}_{AB}^{(2)})},$$

and the correct rate when A lags B is

$$\frac{l(C_{AB}^{(3)} \cap \hat{C}_{AB}^{(3)})}{l(\hat{C}_{AB}^{(3)})}.$$

Table 1 summarizes the simulation results of the correct rates for different α_0 based on 2000 replications, where the parameters $(\alpha, \beta, \alpha_1, \beta_1) = (0.6, 5, 0.6, 0.3)$

are kept fixed. Note that the unconditional variance of ε_t (defined by (2.3))

$$\sigma_0^2 = \mathrm{E}(\varepsilon_t^2) = \frac{\alpha_0}{1-\alpha_1-\beta_1} \tag{3.3}$$

increases as either of the parameters α_0, α_1 or β_1 increases. In Table 1, the second column gives the ratios of σ_0 to $\sigma_A\sqrt{\Delta}$ (the conditional standard deviation of the log price of the stock A, cf. (2.1)), which represents the noise size. The correct rate increases as $\sigma_0/\sigma_A\sqrt{\Delta}$ decreases. Moreover, for fixed $\alpha_0 \le 1.2\times 10^{-8}$, the correct rates of the LCS method are about the same in the fourth to the sixth columns in Table 1. This suggests that the performance of the LCS method is not affected by the receiving order of the information. Nevertheless, the correct rates of the LCS method (cf. column seven in Table 1) can still be improved. In next section, a new pattern matching scheme is proposed to improve the correct rates of the matching positions found by the LCS method.

4. Spectral Analysis of the Bollinger Percents

The real market stock price processes are well recognized as non-stationary processes. One can apply the Bollinger Band to convert a price sequence into a stationary $\%b$ sequence, see for example Wu, Salzberg and Zhang (2004). Bollinger Bands are created by John Bollinger in the early 1980s and are widely used as financial relative high or low indicators of the price. The Bollinger Percent ($\%b$) is obtained from the Bollinger Bands and can be used to measure the highness or lowness of the price relative to previous trades. The bands are curves drawn above and below a simple moving average of period p (the typical value for the period p is 20) by a measure of standard deviation. The three curves are defined as follows:

Middle Bollinger Band = the p-period simple moving average,
Upper Bollinger Band = Middle Bollinger Band + 2 × p-period standard deviation,
Lower Bollinger Band = Middle Bollinger Band − 2 × p-period standard deviation.

The formula for $\%b$ is defined by

$$\%b = \frac{\text{Last Price} - \text{Lower Bollinger Band}}{\text{Upper Bollinger Band} - \text{Lower Bollinger Band}}.$$

Figure 3 is an illustration of Bollinger Bands and $\%b$ of a stock price sequence.

Next, we compute the power spectrum of the $\%b$ sequence. The power spectrum of a stationary sequence decomposes the sequence into a sum of fluctuating components from low to high frequencies. The low-frequency power spectra represent the longer-term trend of the original sequence and the high-frequency power spectra characterize the shorter-time oscillation and the noise. Therefore, we use the low-frequency power spectra of the $\%b$ sequence to acquire its trend. In the following, we excerpt the definition of the power spectrum described in Jones and Pevzner (2004).

Suppose that the complex exponential functions are defined on a finite number of n points, that is, for $t = 1, 2, \cdots, n$. For $\lfloor -n/2 \rfloor + 1 \le k \le \lfloor n/2 \rfloor$, where $\lfloor x \rfloor$ is the floor function of x, the system

$$\left\{ e^{i2\pi kt/n} : \left\lfloor -\frac{n}{2} \right\rfloor + 1 \le k \le \left\lfloor \frac{n}{2} \right\rfloor \right\}, \tag{4.1}$$

contains exactly n functions. The system (4.1) is actually a collection of orthogonal functions. Let $Z_1, Z_2, \cdots, Z_n$ be a sequence of n numbers. This sequence can be regarded as a set of coordinates of a point in an n-dimensional space. And it can be written as a linear combination of the elements of the basis. For a given n-dimensional space, it is known that any set of n orthogonal vectors forms a basis, hence the given sequence, $\{Z_t\}_{t=1}^{n}$, can be written as a linear combination of the orthogonal complex exponential functions given in (4.1). That is,

$$Z_t = \sum_{k=\lfloor -n/2 \rfloor + 1}^{\lfloor n/2 \rfloor} c_k e^{i2\pi kt/n}, \tag{4.2}$$

where

$$c_k = \frac{1}{n} \sum_{t=1}^{n} Z_t e^{-i2\pi kt/n}. \tag{4.3}$$

(4.2) is called the Fourier series of the sequence Z_t and c_k is called the Fourier coefficients. The coefficient $c_0 = \sum_{t=1}^{n} Z_t/n$ is the average value of the sequence. In the following, we denote $2\pi k/n$ by $\omega_k,\ k = 0, 1, \cdots, \lfloor n/2 \rfloor$. These frequencies are called the Fourier frequencies.

For a given periodic sequence Z_t of period n, the energy associated with the sequence in one period is defined as $\sum_{t=1}^{n} Z_t^2$. Multiplying Z_t on the both sides of (4.2), summing from $t = 1$ to $t = n$, and using the relation (4.3), we have

$$\sum_{t=1}^{n} Z_t^2 = n \sum_{k=\lfloor -n/2 \rfloor + 1}^{\lfloor n/2 \rfloor} |c_k|^2, \tag{4.4}$$

where $|c_k|^2 = c_k \bar{c_k}$. (4.4) is known as Parseval's relation for Fourier series. By (4.4), the total energy of a periodic sequence over the whole time horizon $t = 0, \pm 1, \pm 2, \cdots$ is infinite. Hence, we only consider its energy per unit time, which is called the power of the sequence. This is given by

$$\text{Power} = \sum_{k=\lfloor -n/2 \rfloor + 1}^{\lfloor n/2 \rfloor} |c_k|^2.$$

As noted above, the jth harmonic components include the terms for both $k = j$ and $k = -j$ as they correspond to the same frequency $j(2\pi/n)$. Therefore, we can interpret the quantity

$$\begin{cases} f_0 = c_0^2, \\ f_k = |c_{-k}|^2 + |c_k|^2 = 2|c_k|^2, \ k \neq 0, \end{cases} \tag{4.5}$$

from the term in the Fourier series of Z_t at the kth frequency $\omega_k = 2\pi k/n$ as the contribution to the total power. The quantity f_k is called the power spectrum and describes how the total power is distributed over the various frequency components of the sequence $\{Z_t\}_{t=1}^n$.

By using the first m low-frequency power spectra of a stationary sequence, one can obtain a smooth line for describing the dynamic trend of the sequence. For example, Figure 4(a) is the time plot of a $\%b$ sequence and Figure 4(b) is the corresponding trend estimate based on the first 10 low-frequency power spectra. The smooth line in Figure 4(b) mimics the trend of the $\%b$ sequence.

Next we employ the Pearson's chi-square test to access similarity of the trends of the two sources. The procedure is explained below. Reweight the m lowest-frequency power spectra f_k (see (4.5)) by the following:

$$f'_k = \frac{f_k}{\sum_{i=1}^m f_i}, \ k = 1, 2, \cdots, m. \tag{4.6}$$

Since $\sum_{i=1}^m f'_i = 1$, the reweighted power spectrum, $\{f'_k\}_{k=1}^m$ can be viewed as a probability mass function.

We apply the Pearson's chi-square test to test whether the spectrum distributions of two sequences are the same. We regard nf'_k as the number of observations in class k (corresponding to the k-th lowest frequency), for $k = 1, 2, \cdots, m$. In practice, when applying the Pearson's chi-square test, we need :

1. None of the expected number of observations are less than 1;

2. No more than 20% classes are smaller than 5.

If some of $\{nf'_k\}_{k=1}^m$ do not satisfy the above rules, then we merge the m classes into $m'(m' \leq m)$ classes to satisfy the rules and denote the number in these new classes by $\{nf^*_k\}_{k=1}^{m'}$, where f^*_k is the adjusted spectra after merging.

Let $f^*_{A,i}$ and $f^*_{B,i}$ denote the adjusted spectra after merging of Stock A and Stock B respectively. Consider the following hypothesis testing problem:

$$H_0: \ f^*_{A,i} = f^*_{B,i}, \ \ i = 1, \cdots, m',$$

versus the alternative

$$H_1: \ f^*_{A,i} \neq f^*_{B,i}, \ \text{ for some } i.$$

The Pearson's chi-square test statistic is defined as

$$\sum_{i=1}^{m'} \frac{(nf^*_{B,i} - nf^*_{A,i})^2}{nf^*_{A,i}}, \tag{4.7}$$

which has approximately a chi-square distribution with $m' - 1$ degrees of freedom.

5. Simulation and Empirical Studies

We perform simulation study to investigate the performance of the proposed method in detecting the co-movement period of the two price sequences of Stocks A and B. Simulation results based on 2000 replications are presented in Tables 2-6 for different parameter settings. In the tables, "TA" signifies the ratio of "True and Accept", which means that the matching positions found by the LCS method are correct and the chi-square test also accepts the null hypothesis. "FR" stands for the ratio of "False and Reject", which means that the matching positions found by the LCS method are incorrect and the chi-square test rejects H_0. Similarly, "TR" and "FA" are short for the situations of "True and Reject" and "False and Accept", respectively. Hence, "TA+FR" is the correct rate of the proposed method in choosing the co-movement period of two price sequences and "TR+FA" is the error rate. If the "TA+FR" ratio is close to one, then the proposed method significantly improves the accuracy of the LCS method for the co-movement detection problem.

Similar to the results in Table 1, the correct rates of the matching positions found by the LCS method are still not persuasive, especially when the noise effect increases. Recall from (3.3), the standard deviation σ_0 of the noise increases as either of the parameters α_0, α_1 or β_1 increases. The correct rates of the LCS method also decrease (see Tables 2-4) when either of the parameters α_0, α_1 or β_1 increases. Similarly when the parameter α or β decreases, the impact of the noise term also increases, and the performance of the LCS method becomes worse (see Tables 5-6). However, when we apply the Pearson's chi-square test to the adjusted spectra of the Bollinger Percent, significant improvements are achieved. The ratios of "TA+FR" are all greater than 0.980 in the tables. The results indicate the integration of the LCS method and the proposed scheme introduced in the previous section can effectively detect the co-movement periods of the two price sequences.

For the real data application, we consider the intra-daily high frequency stock price data of Bank of America Corporation (BAC) and Bank of New York Mellon Corporation (BK) in June 12, 2002. We divide the normal trading hours into 35 nonoverlapping time periods, $I_1, \cdots, I_{35}$, each with length 11 minutes, and obtain the integrated volatilities in each I_i for the two stocks, denoted by $\{v_i^{BAC}\}_{i=1}^{35}$ and

$\{v_i^{BK}\}_{i=1}^{35}$, respectively. The two categorical sequences, $\{s_i^{BAC}\}_{i=1}^{35}$ and $\{s_j^{BK}\}_{j=1}^{35}$, are obtained by the criterion (3.1). The LCS method is used to find the matching time periods of these two companies and 20 matching pairs are found and listed below:

$$(I_2, I_2), (I_4, I_4), (I_6, I_5), (I_7, I_7), (I_{10}, I_8), (I_{13}, I_9), (I_{14}, I_{12}), (I_{15}, I_{16}),$$

$$(I_{18}, I_{17}), (I_{19}, I_{19}), (I_{20}, I_{21}), (I_{22}, I_{22}), (I_{23}, I_{23}), (I_{24}, I_{25}), (I_{26}, I_{24}),$$

$$(I_{27}, I_{27}), (I_{28}, I_{28}), (I_{33}, I_{29}), (I_{34}, I_{33}), (I_{35}, I_{30}).$$

Next, we use the proposed method in the previous section to examine the co-movement in the intervals of these 20 matching pairs. There are only two matching pairs (I_{19}, I_{19}) and (I_{23}, I_{23}) are concluded to have similar co-movement pattern. Figure 5 plots the price movements within these two matching pairs which also show great similarity visually.

6. Conclusion

This study proposes a scheme to detect the co-movement periods of two stock price processes. The proposed scheme includes 4 steps: (1) Apply the LCS method to find the matching position of the original sequences; (2) For each matching pair, convert the nonstationary price processes to the Bollinger Percent b sequences; (3) Compute the low-frequency power spectra of the $\%b$ sequences to characterize the dynamic trends; (4) Employ Pearson's chi-square test to assess the similarity of the two spectrum distributions. Simulation and empirical studies show that the proposed scheme can effectively detect the co-movement periods of the price sequences. In the future, we will extend the results of this study to develop financial trading strategies or arbitrage strategies when the similar price movements occur.

Acknowledgements

The authors acknowledge helpful comments by two anonymous referees. The research of the first author was supported by the grant NSC 100-2118-M-110-003 from the National Science Council of Taiwan. The research of the third author was supported by the grant NSC 100-2118-M-390-001 from the National Science Council of Taiwan.

Appendix

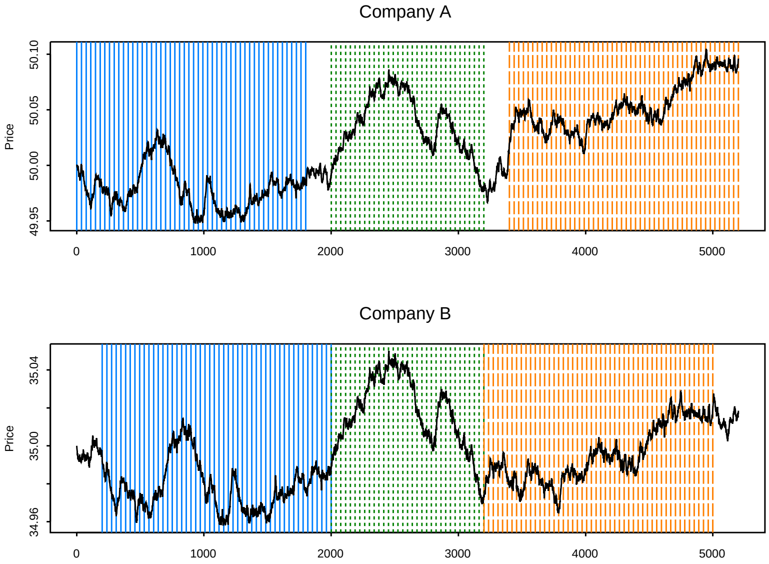


Figure 1: Time plots of the stock prices of companies A and B. The blue region indicates the reaction time of company A leads B 15 minutes. The green region indicates the contemporaneous reactions of companies A and B. And the orange region indicates company A lags B 15 minutes

							w				
			1	2	3	4	5	6	7	8	9
			c	b	a	c	b	a	a	b	a
	1	a	0	0	1	1	1	1	1	1	1
	2	b	0	1	1	1	2	2	2	2	2
	3	c	1	1	1	2	2	2	2	2	2
v	4	d	1	1	1	2	2	2	2	2	2
	5	b	1	2	2	2	3	3	3	3	3
	6	b	1	2	2	2	3	3	3	4	4

Figure 2: The matching positions of the LCSs of string **v** and **w**. These double circles point out the matching positions and the corresponding LCS length

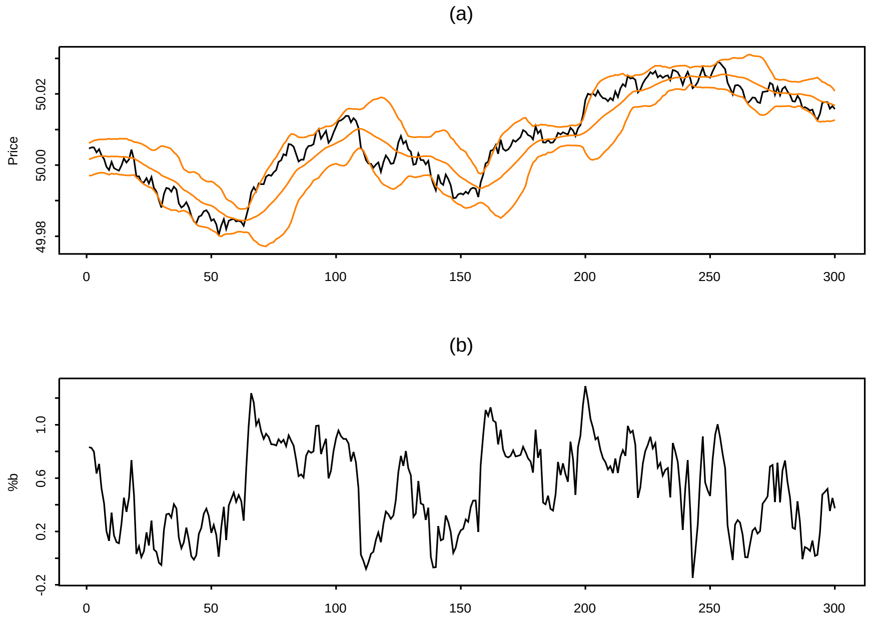
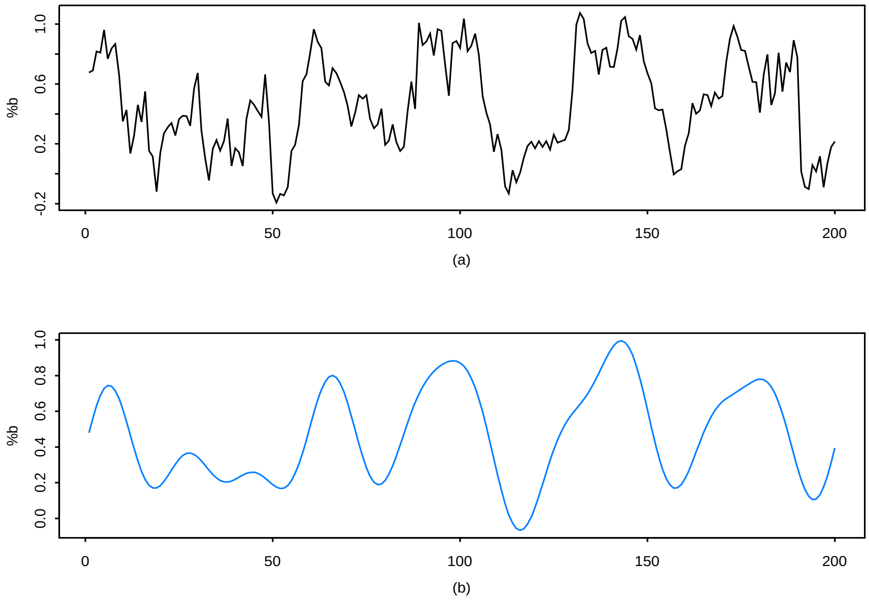


Figure 3: (a) Raw price sequence with Bollinger Bands; (b) The corresponding %b sequence



(a)

(b)

Figure 4: (a) Time plots for a %b sequence; (b) The corresponding trend of the %b sequence computed by the first 10 low-frequency power spectra

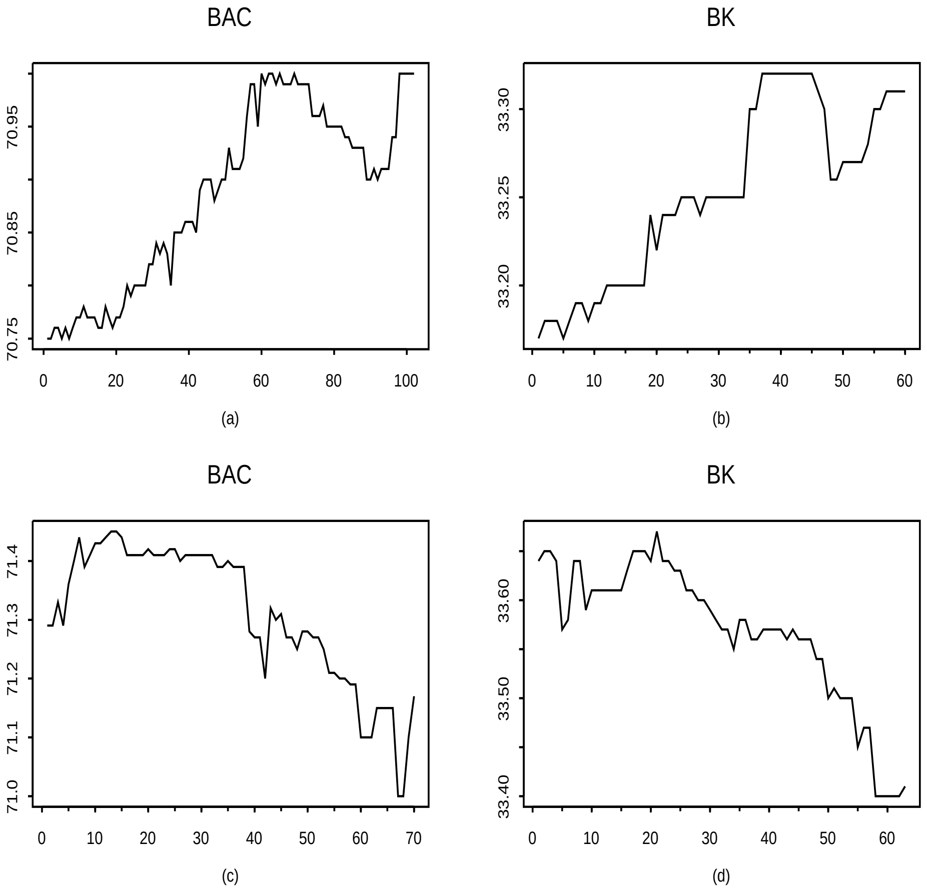


Figure 5: Time-series plots for (a) price data in I_{19} of BAC; (b) price data in I_{19} of BK; (c) price data in I_{23} of BAC; (d) price data in I_{23} of BK

Table 1: The LCS results of simulation data in different α_0 and fixed $\alpha = 0.6$, $\beta = 5$, $\alpha_1 = 0.6$, and $\beta_1 = 0.3$

α_0	$\frac{\sigma_0}{\sigma_A\sqrt{\Delta}}$	$l(\hat{C}_{AB})$	$\frac{l(C_{AB}^{(1)} \cap \hat{C}_{AB}^{(1)})}{l(\hat{C}_{AB}^{(1)})}$	$\frac{l(C_{AB}^{(2)} \cap \hat{C}_{AB}^{(2)})}{l(\hat{C}_{AB}^{(2)})}$	$\frac{l(C_{AB}^{(3)} \cap \hat{C}_{AB}^{(3)})}{l(\hat{C}_{AB}^{(3)})}$	$\frac{l(CMP)}{l(\hat{C}_{AB})}$
1.2×10^{-7}	31.225	9.212	0.113	0.091	0.000	0.049
6.0×10^{-8}	22.079	9.841	0.169	0.277	0.010	0.106
3.0×10^{-8}	15.613	11.537	0.291	0.472	0.138	0.261
1.2×10^{-8}	9.874	14.182	0.508	0.560	0.434	0.491
6.0×10^{-9}	6.982	15.659	0.619	0.606	0.580	0.600
3.0×10^{-9}	4.937	16.431	0.680	0.651	0.665	0.667
1.2×10^{-9}	3.123	17.044	0.719	0.683	0.711	0.707

Table 2: Simulation results of detecting the dynamic co-movement of two stock price sequences by the LCS and the proposed methods with various α_0 and fixed $\alpha = 0.6$, $\beta = 5$, $\alpha_1 = 0.6$, and $\beta_1 = 0.3$

	α_0						
	1.2×10^{-7}	6.0×10^{-8}	3.0×10^{-8}	1.2×10^{-8}	6.0×10^{-9}	3.0×10^{-9}	1.2×10^{-9}
$\frac{l(CMP)}{l(\hat{C}_{AB})}$	0.050	0.099	0.260	0.505	0.600	0.669	0.710
significant level of chi-square test: 0.05							
TA	0.046	0.097	0.255	0.495	0.587	0.655	0.696
FR	0.948	0.899	0.738	0.494	0.400	0.330	0.290
TA+FR	0.994	0.996	0.994	0.988	0.986	0.986	0.986
TR	0.004	0.002	0.005	0.011	0.013	0.014	0.014
FA	0.002	0.002	0.001	0.001	0.000	0.001	0.000
TR+FA	0.006	0.004	0.006	0.012	0.014	0.014	0.014
significant level of chi-square test: 0.01							
TA	0.048	0.098	0.257	0.497	0.589	0.658	0.698
FR	0.946	0.897	0.736	0.492	0.398	0.329	0.289
TA+FR	0.994	0.995	0.993	0.990	0.988	0.988	0.987
TR	0.002	0.001	0.004	0.008	0.011	0.011	0.011
FA	0.004	0.004	0.003	0.002	0.002	0.001	0.001
TR+FA	0.006	0.005	0.007	0.010	0.012	0.012	0.013

Table 3: Simulation results of detecting the dynamic co-movement of two stock price sequences by the LCS and the proposed methods with various α_1 and fixed $\alpha = 0.6$, $\beta = 5$, $\alpha_0 = 1.2 \times 10^{-8}$, and $\beta_1 = 0.3$

	α_1					
	0.6	0.5	0.4	0.3	0.2	0.1
$\frac{l(CMP)}{l(\hat{C}_{AB})}$	0.505	0.608	0.681	0.698	0.710	0.723
significant level of chi-square test: 0.05						
TA	0.495	0.596	0.666	0.685	0.696	0.709
FR	0.494	0.392	0.319	0.301	0.289	0.276
TA+FR	0.988	0.987	0.985	0.987	0.985	0.985
TR	0.011	0.012	0.014	0.013	0.015	0.014
FA	0.001	0.001	0.001	0.000	0.000	0.001
TR+FA	0.012	0.013	0.015	0.013	0.015	0.015
significant level of chi-square test: 0.01						
TA	0.497	0.598	0.669	0.687	0.698	0.712
FR	0.492	0.391	0.318	0.300	0.289	0.276
TA+FR	0.990	0.989	0.987	0.988	0.987	0.987
TR	0.008	0.010	0.012	0.011	0.012	0.011
FA	0.002	0.002	0.001	0.001	0.001	0.001
TR+FA	0.010	0.011	0.013	0.012	0.013	0.013

Table 4: Simulation results of detecting the dynamic co-movement of two stock price sequences by the LCS and the proposed methods with various β_1 and fixed $\alpha = 0.6$, $\beta = 5$, $\alpha_0 = 1.2 \times 10^{-8}$, and $\alpha_1 = 0.3$

	β_1					
	0.6	0.5	0.4	0.3	0.2	0.1
$\frac{l(CMP)}{l(\hat{C}_{AB})}$	0.471	0.640	0.684	0.698	0.718	0.719
significant level of chi-square test: 0.05						
TA	0.462	0.627	0.670	0.685	0.705	0.706
FR	0.528	0.359	0.316	0.301	0.282	0.280
TA+FR	0.990	0.985	0.986	0.987	0.987	0.986
TR	0.009	0.014	0.014	0.013	0.013	0.013
FA	0.001	0.001	0.001	0.000	0.000	0.000
TR+FA	0.010	0.015	0.014	0.013	0.013	0.014
significant level of chi-square test: 0.01						
TA	0.464	0.629	0.672	0.687	0.707	0.708
FR	0.526	0.358	0.315	0.300	0.281	0.280
TA+FR	0.990	0.987	0.987	0.988	0.988	0.988
TR	0.008	0.011	0.011	0.011	0.010	0.011
FA	0.002	0.002	0.001	0.001	0.001	0.001
TR+FA	0.010	0.013	0.013	0.012	0.012	0.012

Table 5: Simulation results of detecting the dynamic co-movement of two stock price sequences by the LCS and the proposed methods with various α and fixed $\beta = 5$, $\alpha_0 = 1.2 \times 10^{-8}$, $\alpha_1 = 0.6$, and $\beta_1 = 0.3$

	α						
	0.8	0.7	0.6	0.5	0.4	0.3	0.2
$\frac{l(CMP)}{l(\hat{C}_{AB})}$	0.562	0.536	0.505	0.436	0.379	0.297	0.186
significant level of chi-square test: 0.05							
TA	0.553	0.525	0.495	0.426	0.369	0.287	0.175
FR	0.437	0.463	0.494	0.562	0.620	0.702	0.812
TA+FR	0.990	0.989	0.988	0.988	0.989	0.989	0.986
TR	0.009	0.010	0.011	0.011	0.010	0.009	0.012
FA	0.001	0.001	0.001	0.001	0.001	0.002	0.002
TR+FA	0.010	0.011	0.012	0.012	0.011	0.011	0.014
significant level of chi-square test: 0.01							
TA	0.555	0.527	0.497	0.428	0.372	0.290	0.180
FR	0.436	0.462	0.492	0.561	0.618	0.700	0.809
TA+FR	0.991	0.989	0.990	0.989	0.990	0.989	0.990
TR	0.007	0.008	0.008	0.009	0.007	0.007	0.006
FA	0.002	0.002	0.002	0.003	0.003	0.004	0.004
TR+FA	0.009	0.011	0.010	0.011	0.010	0.011	0.010

Table 6: Simulation results of detecting the dynamic co-movement of two stock price sequences by the LCS and the proposed methods with various β and fixed $\alpha = 0.6$, $\alpha_0 = 1.2 \times 10^{-8}$, $\alpha_1 = 0.6$, and $\beta_1 = 0.3$

	β					
	1	5	10	15	20	25
$\frac{l(CMP)}{l(\hat{C}_{AB})}$	0.254	0.505	0.561	0.581	0.595	0.610
significant level of chi-square test: 0.05						
TA	0.250	0.495	0.544	0.563	0.576	0.592
FR	0.745	0.494	0.438	0.418	0.404	0.389
TA+FR	0.995	0.988	0.983	0.982	0.980	0.981
TR	0.003	0.011	0.016	0.018	0.019	0.018
FA	0.002	0.001	0.001	0.001	0.001	0.001
TR+FA	0.005	0.012	0.017	0.018	0.020	0.019
significant level of chi-square test: 0.01						
TA	0.251	0.497	0.548	0.566	0.580	0.595
FR	0.743	0.492	0.437	0.417	0.403	0.388
TA+FR	0.994	0.990	0.985	0.984	0.983	0.983
TR	0.003	0.008	0.013	0.015	0.015	0.016
FA	0.003	0.002	0.002	0.002	0.002	0.002
TR+FA	0.006	0.010	0.015	0.016	0.017	0.017

References

Agrawal, R., Faloutsos, C. and Swami, A. (1993). Efficient similarity search in sequence databases. In *Proceedings of 4th International Conference on Foundations of Data Organization and Algorithms* **730**, 69-84. Chicago.

Bergroth, L., Hakonen, H., and Raita, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings of the 7th International Symposium on String Processing Information Retrieval* (*SPIRE*), 39-48. IEEE Computer Society.

Dacorogna, M. M., Gençay, R., Müller, U. A., Olsen, R. B. and Pectet, O. V. (2001). *An Introduction to High-Frequency Finance.* Academic Press, San Diego, California.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM* **18**, 341-343.

Jones, N. C. and Pevzner, P. A. (2004). *An Introduction to Bioinformatics Algorithms.* MIT Press, Cambridge, Massachusetts.

Wu, H., Salzberg, B. and Zhang, D. (2004). Online event-driven subsequence matching over financial data streams. In *proceeding of Special Interest Group on Management of Data* (*SIGMOD*), 23-34.

Received September 13, 2011; accepted December 14, 2011.

Guo, Mei-Hui
Department of Applied Mathematics
National Sun Yat-sen University
70, Lienhai Rd., Kaohsiung 80424, Taiwan
guomh@math.nsysu.edu.tw

Liu, Ching-An
Department of Applied Mathematics
National Sun Yat-sen University
70, Lienhai Rd., Kaohsiung 80424, Taiwan

Huang, Shih-Feng
Department of Applied Mathematics
National University of Kaohsiung
700, Kaohsiung University Rd., Kaohsiung 81148, Taiwan
huangsf@nuk.edu.tw