

A Data mining algorithm to analyse stock market data using lagged correlation.

Cicil Fonseka

School of Computing and Mathematics
University of Western Sydney
Campbelltown, Australia
cfons@doh.health.nsw.gov.au

Liwan Liyanage

School of Computing and Mathematics
University of Western Sydney
Campbelltown, Australia
l.liyanage@UWS.edu.au

Abstract - This paper develops an algorithm for predicting the market direction more accurately when two stocks are strongly correlated to each other with a lag of K number of trading days. The forecasting horizon is the lag; therefore this method is suitable for short term capital gains when the correlation is strong. This will identify the stocks that are closely related, display the daily price movements and its direction side by side and forecast the direction of the price movement for the dependent stock as well as clearly showing the applicable lag. To test the effectiveness of the method, the most correlated stocks were found and prediction of the direction of the price movements made for 3 different dates for training the model. For each date actual data were then used to verify the accuracy of the prediction. In the testing and verification stage the model predicted the direction of the movement of the stock prices accurately 67% of the time.

A generic algorithm is specified so that an automated data mining process can be developed. This algorithm takes into consideration the market-wise analysis performed, varying the lag from a lower limit to an upper limit as specified by the user, calculating the correlation coefficient for each independent stock and all other dependent stocks in the market, selects the pairs of stocks where the correlation coefficients are above a user specified range and lists the stocks data graphically side by side for easy comparison.

The primary motivation of this paper is threefold. First, this research examines and analyses the use of market-wide lagged correlation analysis as a forecasting tool. Specifically the ability of one stock to predict the future usually short term future trends of a closely correlated another stock. Second, this paper endeavours to determine the feasibility and practicality of using lagged correlation analysis as a forecasting tool for the individual investor. Finally this paper specifies the general algorithm for the process so that it can be automated in a data mining technique

In summary, the paper finds ways for the investor to reduce the short term risk of investing in the share market.

Keywords— Data Mining, Lagged Correlation, Stock Market, Predictive Modelling, Stock Market Strategy, Stock Market Algorithm.

I. INTRODUCTION

Individuals, researchers, investors, financial professionals, are continually looking for a superior system which will yield them high returns. One of the best known concepts in finance is that markets are efficient. An efficient market adjusts prices

without delay to reflect all available public information thus making it not possible to make excessive profits. The efficient market hypothesis was associated with the idea of a "random walk". However, financial economists now believe that our securities markets are at least partially predictable.

Since the idea of the random walk research have attempted to find ways to improve the predictability of the market. In Beyond the Random Walk, Singal [1] discussed the concept of market efficiency and anomalies to the Random Walk hypothesis including the frequency of the mispricing, the financial instruments that can be used, and the number of transactions per year. Kim Suk-Joong and McKenzie [2] considered the relationship between stock market autocorrelation, the presence of international investors and the stock market volatility. Chui, Andy C.W. and Kwok, Chuck [3] evaluated the Cross-autocorrelation between A (the shares owned by Chinese citizens only) Shares and B Shares (shares owned by foreigners only) in the Chinese Stock Market. K. Lam and K.C. Lam [4] attempted to improve trading results by forecasting a key summary statistic of future prices using neural network for the generation of trading signals. Bettini, C., Wang, X.S., Jajodia [5] used multiple granularities in time sequences in mining temporal relationships to predict the stock market. David Morena and Ignacio Olmeda [6] found that Artificial Neural Networks do not provide superior performance than the linear models. Chris Brooks [7] emphasizes that and researchers are still uncertain as to the precise role of volume in the analysis of financial markets as a whole. Andrew W. Lo and Jiang Wang [8] found the joint behaviour of price and quantity reveals more information about the relation between asset prices and economic factors than do prices alone.

The primary motivation of this paper is threefold. First, this research examines and analyses the use of market-wide lagged correlation analysis as a forecasting tool. Specifically the ability of one stock to predict the future usually short term future trends of a closely correlated another stock. Second, this paper endeavors to determine the feasibility and practicality of using lagged correlation analysis as a forecasting tool for the individual investor. As a part of this research to achieve this purpose, a software tool has been developed which has a simple graphical user interface. This can be used to perform the required analysis by an investor. Thirdly, accuracy of

results of the model is compared against a traditional forecasting method, linear regression analysis and the probability of the model's forecast being correct is calculated.

Overall, the paper makes three major contributions. First, it identifies the most correlated stocks with a lag K trading days in the Australian Stock Exchange and using the tool an investor can repeatedly generate the information every time the market data are updated. Secondly the investor has a general idea to what extent the correlation provides an edge through the duration of the lag. Finally this paper specifies the general algorithm for the process so that it can be automated in a data mining technique. In summary, the paper finds ways for the investor to reduce the short term risk of investing in the share market.

II. METHODOLOGY

A. Data

The daily Australian Stock Exchange, stock market data were downloaded from <http://www.float.com.au/scgi-bin/prod/dl.cgi> for the period covered by the research. The data for each day comprises of Stock name, Date, Open price, Close price, Daily highest price, Daily lowest price and the volume of stock traded on that day. Considering the high volume of analysis to be performed a software tool was specifically developed as a part of this research. This software tool -Stock Strategy Analyser (SSA) can upload the files in text format, store historical Stock Market data, perform the necessary data transformation and analyse the data according the algorithms developed for this research.

B. Selection of Data

For this analysis daily stock market data were used for all companies which has daily data between 01/07/2005 and 29/07/2008. If a company had started after 01/07/2008 that company was excluded from the analysis for lack of historical data. Since the investment strategy being researched is a short term strategy, if the company was not trading on the 29/07/2008 (the date the analysis was carried out) that company was also excluded.

C. Data Transformation

This research being a market-wide analysis and the emphasis is on the direction of the movement of stock prices it was necessary to perform the correlation analysis on an index of the daily closing price rather than the raw daily price of the stock. The SSA tool was programmed to perform the analysis after transforming the data of each company to an index. For this purpose the Closing price on the 01/07/2005 was considered as the base or 100 and the subsequent prices were expressed as an index to that value. This kept track of the relative movement of prices for all companies.

III. ANALYSIS

A. Correlation Analysis.

Correlation analysis measures the inter-relationship between the stock prices of two companies. SSA can use the stock price of company A as the dependent variable and the stock prices of all other listed companies as the independent variables to calculate correlation coefficients.

SSA allows the user to forward shift periods (lag). The independent variable can be forwarded (i.e. 10 days) to see if there is any predictive power in the indicator. SES can isolate companies which has lagged correlation > 0.95 or < -0.95 (or user input value for r) and chart them side by side.

The correlation coefficients are calculated based on an index of price movements rather than the actual price of the stock. This allows comparison of high priced and low priced stocks possible.

It is possible for a certain security to lead other securities of move in an opposite direction. (Refer to Fig. 1 - Correlation Analysis with Lag.)

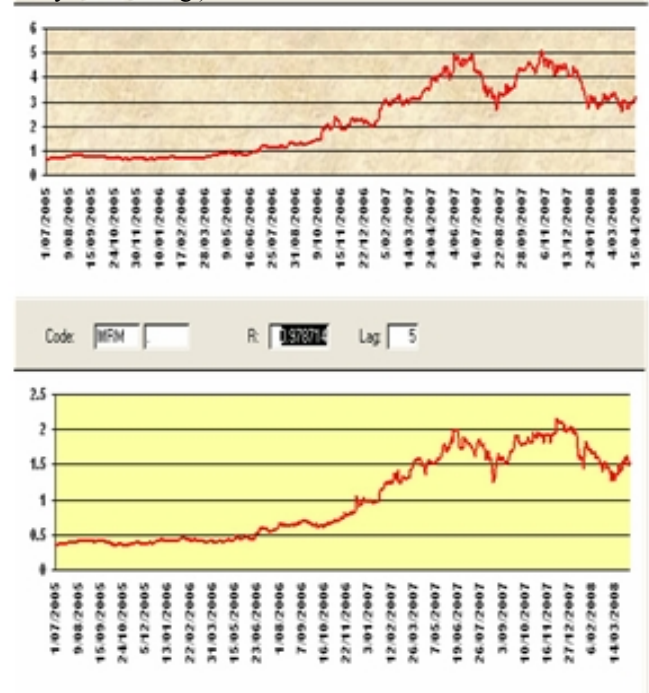


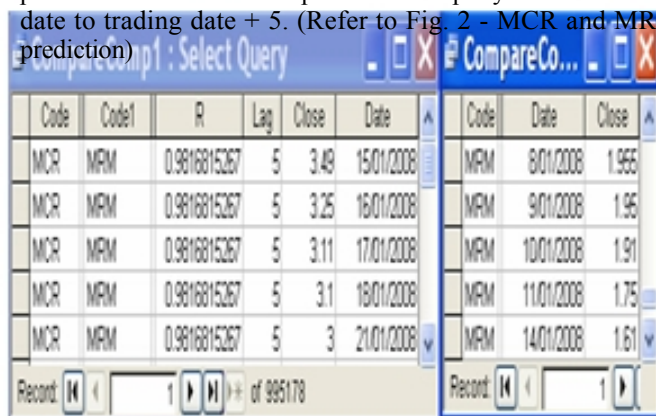
Fig. 1. Correlation analysis with lag

The chart shows the dependent variable security MCR (Minco Resources) which is positively related to the independent variable security of MRM (Mermaid Marine). The correlation coefficient is 0.98 and the lag is 7 days. Because of the strong correlation and the 7-day lag the direction of the price movement of MCR can be used as an indication of the direction of the price movement of MRM.

SSA allows the user to specify a range in number of days (say 1 day to 20 days) to forward shift periods (lag). SES then calculates the correlation coefficients using lags of 1 day to 20 days. The companies which have over a user specified correlation coefficient r , are presented in a graph side by side for comparison. With this method it is possible to find the lag which would provide the strongest correlation. The correlation coefficients are calculated based on an index of price movements rather than the actual price of the stock. This allows comparison of high priced and low priced stocks possible.

B. Forward Testing and Verification

To test the predictability of the model, the marketwise analysis was performed on two cut-off dates -15 Jan 2008 and 15 Feb 2008. The SSA was given data up-to each date in the past and the market-wide correlation analysis was performed. SSA identifies the companies with the highest degree of correlation and sorts them in descending order of r value. Then SSA was provided with the actual data for the companies it is predicting and the accuracy of the prediction is evaluated. What is predicted is the direction of the price movement within the time frame the prediction is applicable. As an example if the index of price movements of company A and B are correlated with each other with $r > .R$ or $r < -.R$ with a lag k of 5 days on the trading date d , the local trend in prices of company B from trading date -10 to trading date is used to predict the local trend in prices of company A from trading date to trading date + 5. (Refer to Fig. 2 - MCR and MRM prediction)



Code	Code1	R	Lag	Close	Date
MCR	MRM	0.9816815267	5	3.48	15/01/2008
MCR	MRM	0.9816815267	5	3.25	16/01/2008
MCR	MRM	0.9816815267	5	3.11	17/01/2008
MCR	MRM	0.9816815267	5	3.1	18/01/2008
MCR	MRM	0.9816815267	5	3	21/01/2008

Fig. 2 MCR and MRM - price movements

MCR prices are shown on the left. The direction of the price movement in the 5 day period is down. This direction can be used as an indication of the direction of the dependent company MRM with a lag of 5 days. In this case MRM also shows a downward trend in the direction of the price movement. The results are listed in table 1.

TABLE I. PREDICTED AND ACTUAL TREND

Trading Date	Lag	Trading Date+Lag	Company Code A	Company Code B	Predicted Trend	Actual Trend	Long Term Prediction
08/01/08	5	15/01/2008	MCR	MRM	Down	Down	Up
07/01/08	6	15/01/2008	MCR	MRM	Down	Down	Up
04/01/08	7	15/01/2008	MCR	MRM	Down	Down	Up
03/01/08	8	15/01/2008	MCR	MRM	Down	Down	Up
08/01/08	5	15/01/2008	SFY	STW	Down	Down	Up
08/01/08	7	15/01/2008	IRE	MLB	Down	Down	Up
07/01/08	6	15/01/2008	IRE	MLB	Down	Down	Up
07/01/08	6	15/01/2008	DTL	SMX	Down	Down	Up
08/01/08	5	15/01/2008	CPB	PLA	Down	Up	Up
07/01/08	6	15/01/2008	IRE	MLB	Down	Down	Up
04/01/08	7	15/01/2008	DTL	SMX	Up	Down	Up
04/01/08	7	15/01/2008	CPB	PLA	Up	Down	Up
03/01/08	8	15/01/2008	CPB	PLA	Up	Up	Up
08/01/08	5	15/01/2008	AFI	CBA	Down	Down	Up
08/02/08	5	15/02/2008	MCR	MRM	Down	Down	Up
07/02/08	6	15/02/2008	MCR	MRM	Down	Down	Up
06/02/08	7	15/02/2008	MCR	MRM	Down	Down	Up
05/02/08	8	15/02/2008	MCR	MRM	Up	Down	Up
01/02/08	10	15/02/2008	MCR	MRM	Up	Down	Up
08/02/08	5	15/02/2008	SFY	STY	Down	Down	Up
08/02/08	5	15/02/2008	IRE	MLB	Down	Down	Up
08/02/08	5	15/02/2008	DTL	SMX	Down	Down	Up
07/02/08	6	15/02/2008	IRE	MLB	Up	Down	Up
08/02/08	5	15/02/2008	CPB	PLA	Down	Down	Up
07/02/08	6	15/02/2008	CPB	PLA	Down	Down	Up
06/02/08	7	15/02/2008	IRE	MLB	Down	Down	Up
06/02/08	7	15/02/2008	DTL	SMX	Down	Down	Up

TABLE II. PREDICTED AND ACTUAL DIRECTION OF PRICE MOVEMENTS -SUMMARY RESULTS

	Actual Up	Actual Down	Total
Predicted Down	1	20	21
Total	2	25	27

Correctly predicting the downward price trend provides an opportunity for sell signals or put warrants. Predicting correctly the upward movements allows the investor to make a short term capital gain.

C. Generic Algorithm for Automating the Technique for Data Mining

User input parameters are Lag and R
 $m = 0$
 For each $i = 1$ to n (stocks in the market)
 Get all dates and closing price

```

each (j + i) to n (stocks in the market)
    Get all dates and closing price
    For each lag for 1 to K
        Shift the closing price by lag
        Calculate the correlation coefficient between the two stocks
        If  $r > R$  or  $r < -R$  then
             $M = M + 1$ 
            Stock1NameArray(m) = Stock(i)
            Stock1RArray(m) = mod(r)
            Stock2NameArray(m) = Stock(j)
            Stock2RArray(m) = mod(r)
        End if
    Next Lag
Next j
Next i

```

IV. CONCLUSION

This paper developed an algorithm for predicting the market direction more accurately when two stocks are strongly correlated to each other with a lag of K number of trading days. A generic algorithm is specified so that an automated data mining process can be programmed. This will identify the stocks that are closely related, display the daily price movements and its direction side by side and forecast the direction of the price movement for the dependent stock as well as clearly showing the applicable lag which is the forecasting horizon

V. FURTHER RESEARCH

As part of this project it is planned to conduct further research into the impact of unusually high volume of trading and the direction of the stock price.

REFERENCES

- [1] Vijay Singal, "Beyond the Random Walk: A Guide to Stock Market Anomalies and Low-Risk Investing", Oxford University Press, 2004.
- [2] Kim, Suk-Joong and McKenzie, Michael D., "Conditional Autocorrelation and Stock Market Integration" 2006, Available at SSRN: <http://ssrn.com/abstract=943970>
- [3] Chui, Andy C.W. and Kwok, Chuck C.Y., "Cross-autocorrelation between A Shares and B Shares in the Chinese Stock Market." Journal of Financial Research, Sep 1998.
- [4] K. Lam and K.C. Lam. "Forecasting for the generation of trading signals in financial markets" Journal of Forecasting, 19, pp. 39-52. 2000
- [5] Bettini, C., Wang, X.S., Jajodia, S., "Mining temporal relationships with multiple granularities in time sequences.", Data Engineering Bulletin (1998) 21:32-38.
- [6] David Morena and Ignacio Olmeda, "Is the predictability of emerging and developed stock markets really exploitable?" European Journal of Operational Research Volume 182, Issue 1, 1 October 2007, Pages 436-454
- [7] Chris Brooks, "Predicting Stock Index Volatility: Can Market Volume Help?" Journal of Forecasting, Vol. 17, 59-80