

Stock Price Analysis Using Univariate and Multivariate Models

Wenhao Leng

Student ID: 1400015475

Xiaoyu Zhang

Student ID: 1500015902

Zhixuan Shao

Student ID: 1400012141 and

Abstract

This project aims at discovering the relationship between the stock prices and returns of one company (MERCHANTS BANK) cross-listed both in Shanghai and Hong Kong. First we considered two univariate models, the ARMA model and the ARMA-GARCH model, which turned out to perform poorly at forecasting. Then we moved on to VAR model, and saw great improvement, especially in H-share price. Through granger causality test, impulse response and variance decomposition analysis, we concluded that A-share price had remarkable influence in explaining H-share price while a large part of A-share price was implied by SSI. Out of our expectation no significant cointegration relationship on the price level was found for this particular stock. In a nutshell, VAR performs better than univariate models as it makes use of more information available. Still, the stock price is highly stochastic and far from predictable.

Keywords: ARMA, GARCH, VAR, Granger-Causality, Variance Decomposition, Impulse Response, Forecasting Performance

CONTENTS

I. Introduction	3
II. Reviews of related studies	3
III. Data Description and Preprocessing	3
A. Dividing Training Set and Test Set	3
B. Transformation and Stationary Test	4
IV. Univariate Models	4
A. ARMA	4
1. Order Selection: ACF, PACF and EACF	4
2. Model Fitting – MA(1)	4
3. Forecasting Performance	4
4. Model Diagnosis	5
B. Variance Model: ARMA-GARCH	5
1. ARCH Effect: significant	5
2. Model Specifying and Fitting	5
3. Model Diagnosis	5
4. Forecasting Performance	5
V. Multivariate Model: VAR	6
A. VAR Model Fitting: VAR(2)	6
B. Granger Causality	6
C. Forecasting Performance	6
D. Variance Decomposition and Impulse Resoponse Function	6
E. Cointegration Test	6
VI. Conclusions	6
References	7

I. INTRODUCTION

In order to achieve its exchange rate target, Chinese government still holds on to the constraints of free capital flow. As a result, mainland and Hong Kong stock markets are segmented, which means that investors cannot convert the shares of the same company listed in Shanghai to its counterpart in Hong Kong. And it has been a common phenomenon that the stock price at mainland enjoys a premium over Hong Kong for most cross-listed companies, even if they actually share the same right. Therefore, this project is entailed to figure out whether there is some implicit relationship between the two different share prices and their returns.

II. REVIEWS OF RELATED STUDIES

There have been tremendous research over the relationship between A and H share and they mainly cover two major aspects: price difference and cointegration.

Xu (2009) [2] tested the factors that result in the price difference between A and H-share of Chinese dual-listed companies based on dynamic panel data models. The empirical results suggest that Differential Demand Hypothesis, the Liquidity Hypothesis and “Through Train” had a great effect on the price difference.

Some scholars also analyzed the cointegrating relation of A and H-share. Yao (2007) [3] implemented empirical research on the contribution effect of A and H-share market on the price of Chinese dual-listed companies. The result suggests that some stocks’ closing prices have cointegrating relation. Liu and Chen (2013) [4] found that SSI and HIS have long-term stable cointegrating relation. Then they found unidirectional causality between these two markets that the performance of the A-share market significantly guides that of the H-share market based on Granger causality test. They also testify this result using variance decomposition. Yan and Cui (2012) [5] works out price indexes based on the selected 57 A and H shares, and separates these indexes into two phases by the point of the date, March 7, 2011, the day when same time opening started. Based on impulse response and variance decomposition analysis, they found that after the same time opening, the relationship of linkage between the A shares and H shares has been improved. The effect of the H shares price to the A shares price appears more obviously.

III. DATA DESCRIPTION AND PREPROCESSING

The original data consist of four time series:

- a. **A**: Shanghai Stock Index SCI: 000001
- b. **H**: Hangseng Stock Index HSI
- c. **AZ**: MERCHANTS BANK SHA: 600036
- d. **HZ**: MERCHANTS BANK HK: 03968

Each seires is the **daily Open Price** of that stock, from 2007/5/30 to 2018/5/29. Note that some dates of **A** and **H** are different due to different holiday schedules. We deal with this problem by keeping only common dates shared by **A** and **H**. Finally we have 4 time series with the same length 2595.

A. Dividing Training Set and Test Set

In order to fairly judge the forecasting performance of a model, we split the data into training set and test set at the very beginning. We leave the last 100 days’ observations (from 2017/12/20 to 2018/5/29) as the test set.

Such a division is based on the assumption that the structure of our models did not change greatly after 2018.

B. Transformation and Stationary Test

At first glance all 4 series are not stationary. We transformed each series by

$$\mathbf{X} \rightarrow 100 \times \Delta \log \mathbf{X} \quad (1)$$

The differenced log price series is known as the **log return**, approximately the ratio of change to current price. Unsurprisingly, as is the common case for return series, all these "difflog" series now appear stationary. We further conduct the **ADF test** to verify it. All 4 ADF tests turn out to be significant (test statistic far beyond the critical values) so we reject the null hypothesis – unit-root process.

IV. UNIVARIATE MODELS

We focus on **HZ** to illustrate the use of univariate models. Studies of other series are basically the same.

A. ARMA

1. Order Selection: ACF, PACF and EACF

ACF and PACF of **HZ.difflog** are given in Appendix. For lags lower than 20, only the 1st lag ACF and PACF are significant. This is a property of MA(1) process. EACF method also indicates a MA(1) model.

2. Model Fitting – MA(1)

The 'auto.arima()' function in R automatically selects the ARIMA model with the lowest AIC. Here it selects the MA(1) model (with zero mean), the same choice as given by EACF.

The only coefficient of the MA(1) model is significant,

$$\hat{\theta}_1 = -0.0916, \quad \text{se}(\hat{\theta}_1) = 0.0205 \quad (2)$$

and the fitted MA(1) model is: $\hat{\mathbf{H}}\mathbf{Z}_t = e_t + \hat{\theta}_1 e_{t-1}$

3. Forecasting Performance

We try both rolling-window forecast and forecast with fixed model parameters. The results turn out to be almost the same (since the test set is considerably shorter than the training set so refitting does not change the parameters much.) We forecast on day $T + 1$ using information earlier than day T , i.e. we always make **1-step ahead forecast**.

We forecast on the test set (100 days). By comparing with the actual value, we compute the RMSE and out-of-sample R^2 :

$$\text{MSE} = \frac{1}{T} \sum_{i=1}^T (y - \hat{y})^2, \quad \text{RMSE} = \sqrt{\text{MSE}}, \quad R^2 = 1 - \frac{\sum_{i=1}^T (y - \hat{y})^2}{\sum_{i=1}^T (y - \bar{y})^2} \quad (3)$$

Here \mathbf{y} is the series **HZ.difflog** within the test set, with standard deviation 2.444087.

The result of MA(1) is very poor:

$$\text{RMSE} = 2.432285, \quad R^2 = -0.037\% \quad (4)$$

suggesting the MA(1) model useless in forecasting.

4. Model Diagnosis

We reexamine the model by checking the residuals using ACF, PACF and **Ljung Box Test**. No significant serial correlation is shown, i.e. the residual sequence appears to be white noise.

However, there is strong evidence suggesting that it is **not independent white noise**. The most obvious **stylized feature** is **volatility clustering**, which cannot be captured with linear models.

We notice that the residual sequence presents an unusually high volatility around 2008-2009, which can be explained by the financial crisis during that period. There are also some other clusters of high volatility. Since traditional ARMA model assumes constant conditional variance, we instead consider an ARCH (or GARCH) model as following.

B. Variance Model: ARMA-GARCH

1. ARCH Effect: significant

Before setting up a model, we want to make sure whether there is indeed a significant ARCH effect (autoregressive conditional heteroscedasticity).

First, we run the ACF, PACF and Ljung-Box Test of the **squared residual Sequence**, and they all suggest strong serial correlation.

Second, we use the **ARCH-LM (Lagrange Multiplier) test** with lag=3, 6, 12. These tests have significantly small p-value, so we reject H_0 : no ARCH effect.

2. Model Specifying and Fitting

Although the ACF, PACF and AIC may prefer a high-order model like ARCH(8), in practice we do not favor such a complicated model. Instead, a conventional GARCH(1,1) variance model may suffice, which means the volatility h_t satisfies:

$$h_t = \omega + \alpha_1 Z_{t-1}^2 + \beta_1 h_{t-1}, \quad Z_t = \sqrt{h_t} e_t, \quad e_t \sim \text{IID N}(0, 1) \quad (5)$$

For the mean model we still use MA(1), which is also suggested by AIC. Then we estimate the parameters in the MA(1)-GARCH(1,1) model:

$$\text{Mean Model: } \hat{\mu} = 0.0465, \quad \hat{\theta}_1 = -0.0617(**) \quad (6)$$

$$\text{Variance Model: } \hat{\omega} = 0.0460(*), \quad \hat{\alpha}_1 = 0.0575(* * *), \quad \hat{\beta}_1 = 0.9388(* * *) \quad (7)$$

3. Model Diagnosis

Under MA(1)-GARCH(1,1) model, neither the **standardized residuals** nor the **standardized squared residuals** shows significant serial correlation any more, according to the Ljung-Box Test. (see Appendix)

4. Forecasting Performance

$$\text{RMSE} = 2.4294, \quad R^2 = 0.199\% \quad (8)$$

There is some improvement compared with MA(1) model, but still far from satisfactory.

V. MULTIVARIATE MODEL: VAR

We have seen that univariate models all perform poorly at forecasting as the model contains too little information. We then consider using additional factors from other relative time series, by setting up a multivariate model, i.e. a conventional VAR model.

A. VAR Model Fitting: VAR(2)

Different information criteria proposed quite different orders (AIC suggests 10 while SC suggests 1). VAR model with a high order ($p > 3$) will have a large number of parameters, rendering the model too complicated. We make a compromise by choosing $p = 2$, the smallest order that can eliminate serial correlation in residual sequence.

$$\mathbf{X}_t = \mathbf{c} + \mathbf{A}_1 \mathbf{X}_{t-1} + \mathbf{A}_2 \mathbf{X}_{t-2} + \boldsymbol{\epsilon}_t \quad (9)$$

Here we set $\mathbf{c} = \mathbf{0}$. Parameter estimates of $\hat{\mathbf{A}}$ are given in Appendix.

We also applied **model restriction**, namely we refitted the model by leaving out insignificant terms. Ljung-Box test shows that the serial correlation of the residuals is not significant.

B. Granger Causality

The result (see Appendix) shows what we expected: While \mathbf{A} and \mathbf{H} significantly granger-cause \mathbf{AZ} and \mathbf{HZ} , the reverse is not true. Individual Stock does not granger-cause Index price, which is in line with our common sense.

C. Forecasting Performance

$$\text{RMSE} = 2.406403, \quad R^2 = 2.080655\% \quad (10)$$

We see a great improvement compared with univariate models.

D. Variance Decomposition and Impulse Resoponse Function

See results in Appendix. \mathbf{A} and \mathbf{AZ} explained a remarkably large proportion of \mathbf{AZ} 's future move. The variance of forecast of \mathbf{HZ} depends mainly on itself, but also on \mathbf{A} , \mathbf{H} and \mathbf{AZ} . The interpretation of the impulse response plot is somewhat similar to that of the granger causality.

E. Cointegration Test

We highly expect a cointegration relationship between 4 series, since the shapes and trends of the 4 series seem similar. Here we focus on log price instead of differenced log price. We check with ADF test that all 4 log price series are unit-root process. Then we conduct following tests to see if there is any cointegration variable. First, the Johansen procedure (both trace and eigenvalue test) gives $r = 0$, suggests no cointegration variables. Second, the Phillips-Ouliaris test does not suggest significant sign of cointegration either.

VI. CONCLUSIONS

We tried several models to explore the relationships between A-share and H-share prices and returns of one cross-listed company. ARMA and ARMA-GARCH model perform poorly in forecasting while VAR model stands out. Generally speaking, in terms of log return, \mathbf{AZ} and \mathbf{HZ} affect each other reciprocally. \mathbf{AZ} explained around 20% of \mathbf{HZ} 's future move while \mathbf{HZ} seems

to have very little influence on AZ. AZ itself together with A has dominated AZ' s upcoming performance.

- [1] Code and outputs are given in Appendix.
- [2] 徐寿福. “双重上市”公司 A、H 股价格差异的因素研究. 证券市场导报,2009,02;54-60.
- [3] 姚宁. 交叉上市和价格发现: 中国 A 股和 H 股的实证研究. 内蒙古农业大学学报: 社会科学版,2007,01; 63-65.
- [4] 刘燕, 陈勇, 周哲英. A 股、H 股市场协整关系和引导关系实证研究. 金融理论与实践,2013,06;35-38.
- [5] 严宏伟, 崔继刚. 同时开市背景下 AH 股的股价联动性分析. 中国外资,2012,10;228-230