Kent State University

MIS-64060: Fundamentals of Machine Learning

Fall 2021

Stock Forecast and Portfolio Optimization

By:

Jiahao Chen

Background

- Forecast

Stock forecasting refers to the behavior of securities analysts who have a deep understanding of the stock market to predict the future development direction of the stock market and the degree of ups and downs based on the development of the stock market. This predictive behavior is based only on the preconditions established based on assumed factors. In this report, Using the time series model predicts the future price and return. ARIMA model is a good choice for stock prediction because model only needs endogenous variables without other exogenous variables. It can clearly get the relationship between time and stock trends.

The full name of the ARIMA model, the differential integrated moving average autoregressive model, is composed of three parts:

$$ARIMA = AR + I + MA$$

The order of the AR model is represented by the letter p, the order of the MA model is represented by the letter q and d is the number of differences made to make it a stationary series.

- Optimization

The main intention of investors or managers of "securities portfolio" is to build an effective portfolio as much as possible. That is, among the numerous securities in the market, select several securities to combine to obtain the highest return per unit risk level, or the lowest risk per unit return level.

The modern asset portfolio theory is mainly aimed at the possibility of resolving investment risks. "Don't put all your eggs in one basket" is the best metaphor for diversified investment portfolios, and this has become a truth in the modern financial investment world. This article will introduce

and summarize the generation and development of portfolio theory in turn. Various investment portfolio theories and various selection models formed. This report hopes to use the R language to calculate the effective boundary of the Markowitz model and the Sharpe ratio theory and use the Monte Carlo method to calculate the best asset investment ratio scheme based on predicted data.

The basic method for analyzing a portfolio of multiple risk assets is the Markowitz model, which follows 7 basic assumptions:

- Investors follow the principle of maximizing utility.
- The investment period is one period.
- Investors are risk avoiders, that is, under the condition of equal returns, investors choose the lowest risk investment portfolio.
- Investors choose the best investment portfolio based on the mean, variance, and covariance.
- The securities market is complete, without transaction costs, and securities can be subdivided infinitely.
- All funds are used for investment, but short selling is not allowed.
- The correlation coefficient between securities is not -1, there are no risk-free securities, and the expected returns of at least two securities are different.

- The statistics part:

The calculation method of the expected rate of return (expectation) of the investment portfolio is:

$$E(R_p) = \sum_{i=1}^{N} W_i E(R_i)$$

$$R_p: Expected\ returns\ for\ the\ asset\ portfolio$$

$$R_i: Expected\ return\ for\ a\ single\ asset$$

$$W_i: the\ position\ of\ the\ i-th\ asset$$

The risk of a portfolio is generally measured by its variance (or standard deviation):

$$Var(R_p) = \sum_{i=1}^{N} W_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{j \neq 1}^{N} W_i W_j \rho_{ij} \sigma_i \sigma_j$$

$$\sigma: i\ or\ j\ asset\ standard\ deviation$$

$$\rho: Correlation\ coefficient\ between\ asset\ i\ and\ asset\ j$$

Next is to try different investment ratios and draw a graph of these investment ratios. For example:

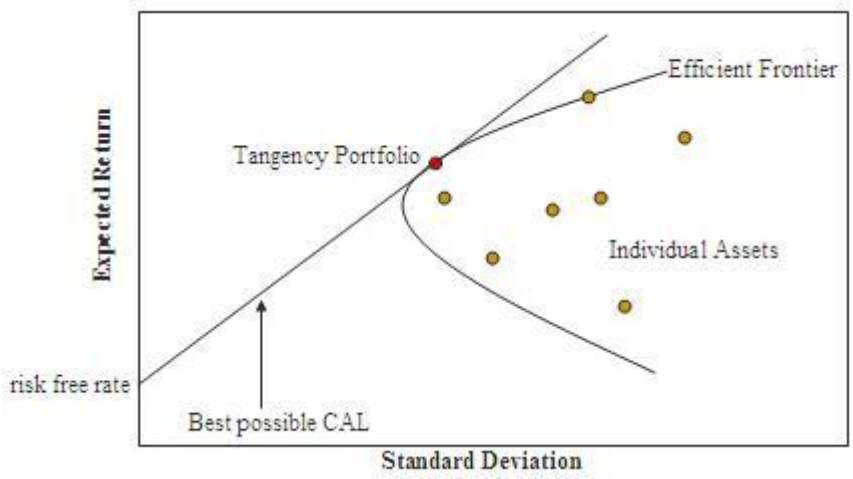


Chart1: Effective Frontier and Sharpe Ratio

We can obtain some effective information to help investment through the effective boundary, such as:

The vertex that Efficient Frontier protrudes to the Y axis is the portfolio with the lowest risk in the portfolio (not the red dot).All points on the asset boundary are the portfolio of investment proportions with the lowest risk under the same expected rate of return.If there are risk-free assets (such as bank financing, treasury bonds, regular (not risk-free, but extremely low risk, which is regarded as risk-free)), they can be invested, and the possibility of investment portfolio will be expanded (Tobin model).When there are risk-free assets, the best investment portfolio (Sharp ratio

theory) is the red dot.

Problem

The problem is constructing a model to predict the future stock price and return of each stock in the portfolio, and calculating what weight of each stock can maximize the return and minimize the risk of the portfolio?

Data description

The source of data is from Yahoo finance (using Quant MoD package in RStudio). I choose 10 stocks to build a portfolio. Here are stock name and price in portfolio.

Stock name	Tickets
Apple Inc.	AAPL
The Home Depot, Inc.	HD
Honeywell International Inc.	HON
Oracle Corporation	ORCL
Johnson & Johnson	JNJ
JPMorgan Chase & Co.	JPM
Microsoft Corporation	MSFT
Amazon.com Inc.	AMZN
Alphabet Inc.	GOOG
Cisco Systems, Inc.	CSCO

Chart 2: Stock Portfolio

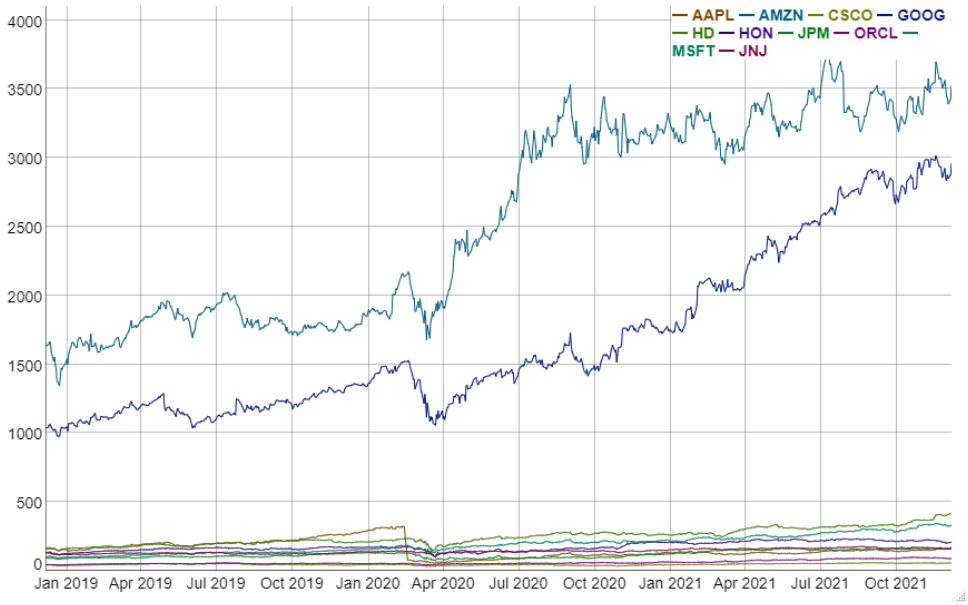


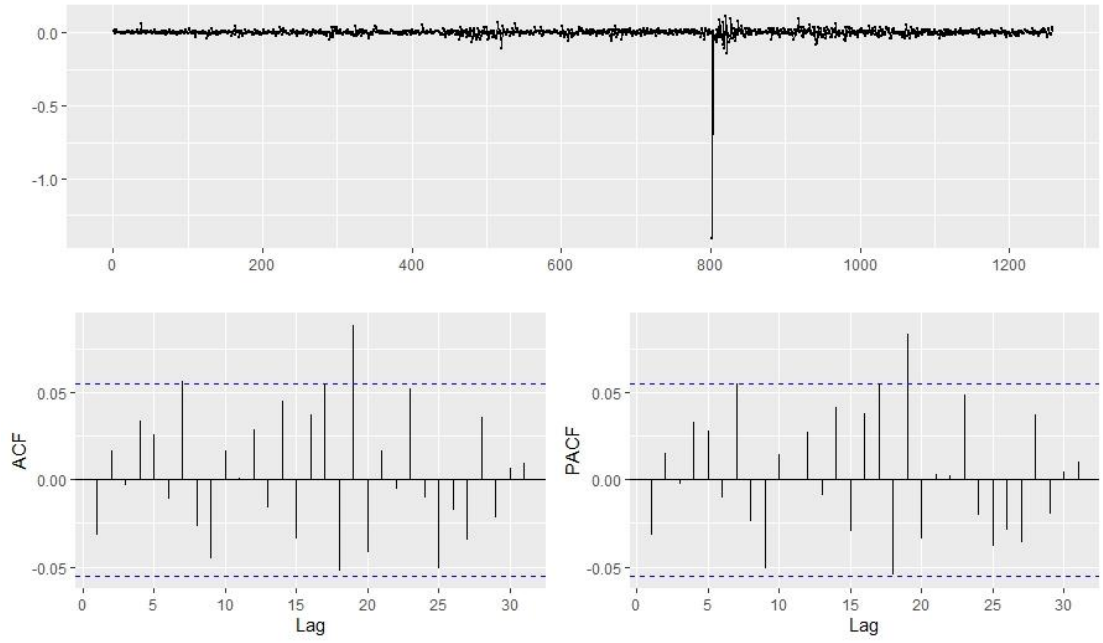
Chart 3: Stock price of Portfolio

Firstly, I use one of stocks in Portfolio (APPL) to create stock prediction model. This is stock price of AAPL.

Chart 4: AAPL Stock Price

As can be seen from Chart 3 and 4, stock price in Portfolio has maintained an upward momentum, except for the US market's diving caused by the Covid-19 epidemic. The ARMA model requires that the sequence participating in the modeling must be stationary. Generally, when dealing with

financial time series, we take the natural logarithm and then differentiate. It also can use in calculating the daily return of each stock. Here are the Apple stock price and portfolio stock price after data processing.




```
        Augmented Dickey-Fuller Test

data:  AAPL_clean
Dickey-Fuller = -8.533, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

Chart 5: Apple stock price after data processing

In Chart 5, by performing ADF test on the sequence after processing, we get its statistical p value<0.01, we reject the null hypothesis and choose the alternative hypothesis: the sequence is stationary.

Chart 6: Portfolio price after data processing

As we can see in chart 6, the general rate of return series fluctuates like this kind of radio waves.

Solution and output

- Prediction

There are two kinds of test to determine the accuracy of the model:

Chart 7: Comparison of ARIMA model forecast results and actual stock price trend (2020.10.21-2021.12.07)

The first test is directly prediction of 300 trading days based on the historical data. As we can see in the Chart 7, We can't predict the accurate stock prices and yields. But the trend in this long period is the same. In other words, if you buy AAPL stocks at OCT 2020, you will gain money in Dec 2021.

To improve the accuracy of the model, we only predict the value at time t+1 each time. When the real closing price at time t+1 is obtained, we will use t+1 day (inclusive) before predicting the closing price at t+2. All previous data (Train Data) to train the model. At the same time, we take the last 300 days in the time series we processed as the test data set (Test Data).

Chart 8: Comparison of ARIMA model forecast results and actual stock price trend (2020.10.21-2021.12.07)

It can be seen from the above results that the final model we obtained satisfactory prediction results.

From the line chart, when the trend is clear (unilateral increase/decrease), the error of the prediction data obtained by the model is small, but the stock price suffers at the turning point, the model cannot predict the fluctuations well, which makes the error larger, which shows that the model has obvious lag in trend changes.

Next is about the error of a single day:

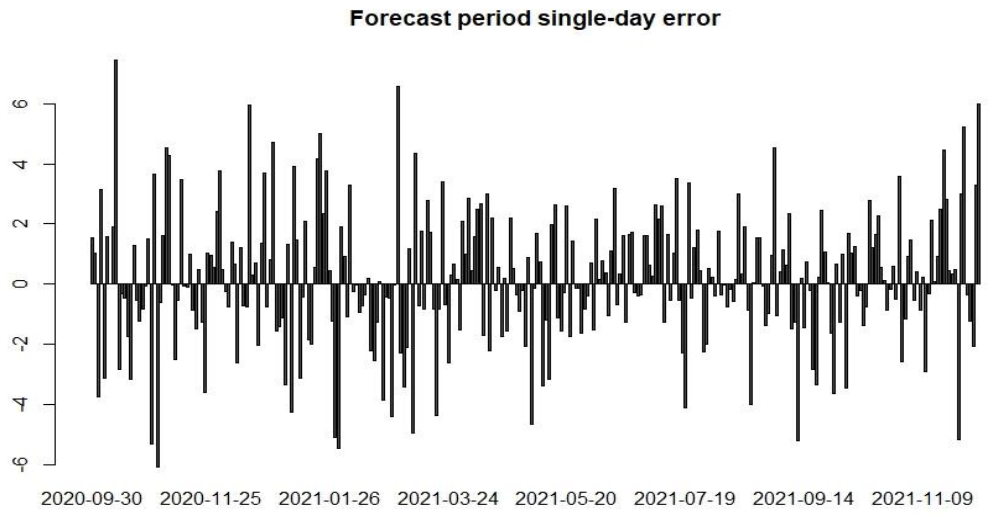


Chart 9: Single-day error in the forecast period (actual value-forecast value)

```
> summary(ts_error)
     Index                error          
 Min.   :2020-09-30   Min.   :-6.07243  
 1st Qu.:2021-01-18   1st Qu.:-1.11676  
 Median :2021-05-05   Median : 0.09079  
 Mean   :2021-05-05   Mean   : 0.13992  
 3rd Qu.:2021-08-20   3rd Qu.: 1.53435  
 Max.   :2021-12-07   Max.   : 7.44468  
```

Chart 10: Statistical data of residual series

From the data and chart, we can see that the average single-day error we got was about 0.13, the median error was 0.09 yuan, and the highest error reached 7.44 single-day. Considering the closing price of the sample stocks reached during the forecast period, I personally think that the statistical error can be said to be controlled within an acceptable range. But then again, as the saying goes, it's

a bit of a mistake, and we have seen the limitations of the ARIMA model used in the article: the inflection point of the stock price cannot be well predicted, and the error increases sharply when the up/down trend changes. With hysteresis and so on. In addition, the ARMA(p,q) that automatically sets the order through the auto. Arima function in this article may be relatively simple, and you can also try to manually set the order to optimize the prediction results.

So, this model is just use for long-term trend prediction and can't predict the accurate stock price.

- Optimization

  In optimization part, I want to use the prices predicted by ARIMA model, but they are not accurate. So, we usually use the historical data to estimate the future expected return and variance. In this model, I use the last 3 years data to calculate the portfolio weights.

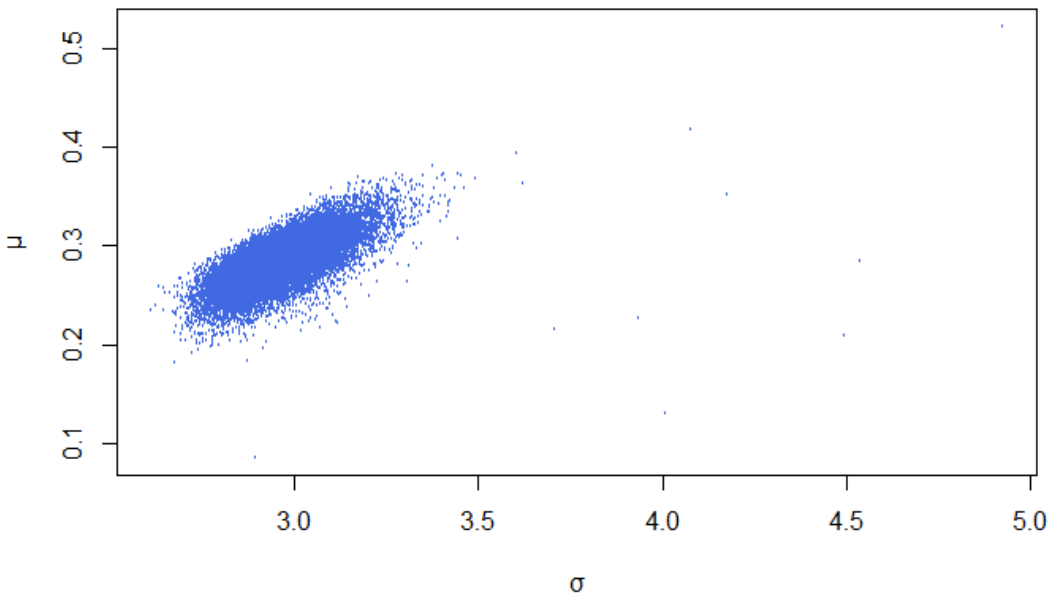


Chart 11: Asset portfolio generated by Monte Carlo method

This chart describes the return rate and standard deviation of a portfolio of 20,000 positions

(investment ratio) randomly generated by the Monte Carlo method. If the portfolio of assets on the left is connected by a curve to form an effective boundary of the portfolio (with lower risk under the same rate of return), the apex of the curve will be the investment solution with the least risk.

To draw this effective boundary, we use the Portfolio library and specify the Treasury bond as our risk-free asset, taking Treasury's 1-day annualized interest rate on December 9th at 0.99%:

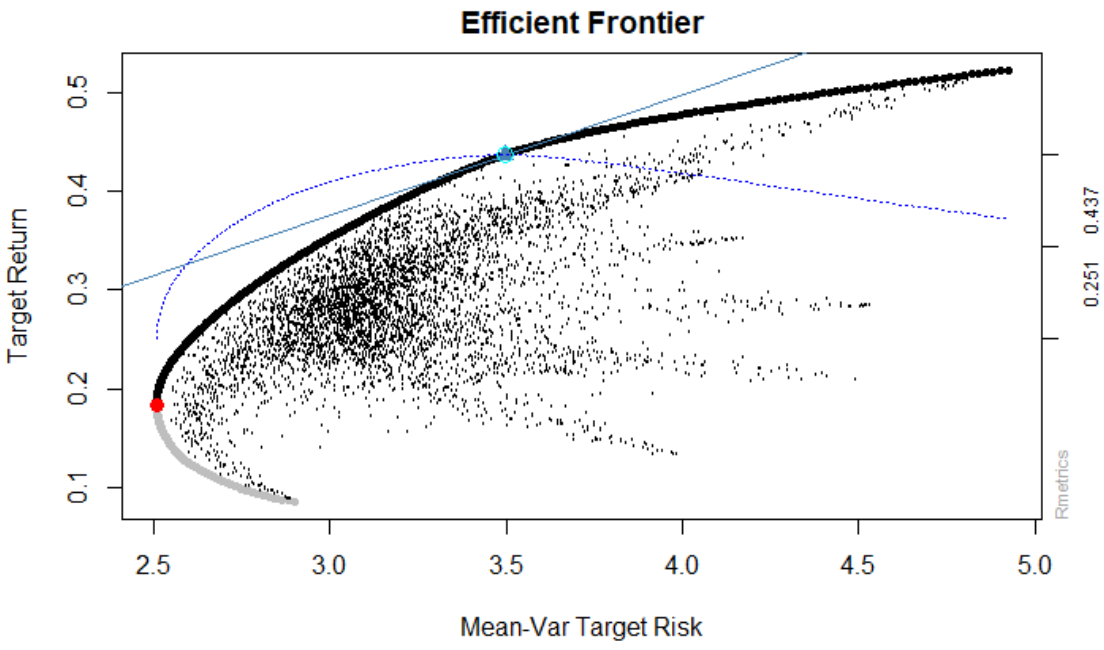


Chart 12: Minimum risk portfolio

The black line represent that the portfolio gets the lowest risk with the target return. In all investment portfolios, the investment portfolio represented by the red dot has the lowest relative risk in the pursuit of the highest rate of return. In all investment portfolios, if there is a risk-free asset with an annual interest rate of 0.99%, the combination represented by the blue triangle is the best investment portfolio under the maximum Sharpe ratio.

Next, to arrive at a suitable position arrangement, we need to find the positions represented by the above two best combination points (combination of investment ratios):

1. The minimum risk portfolio based on the efficient frontier theory (red point):

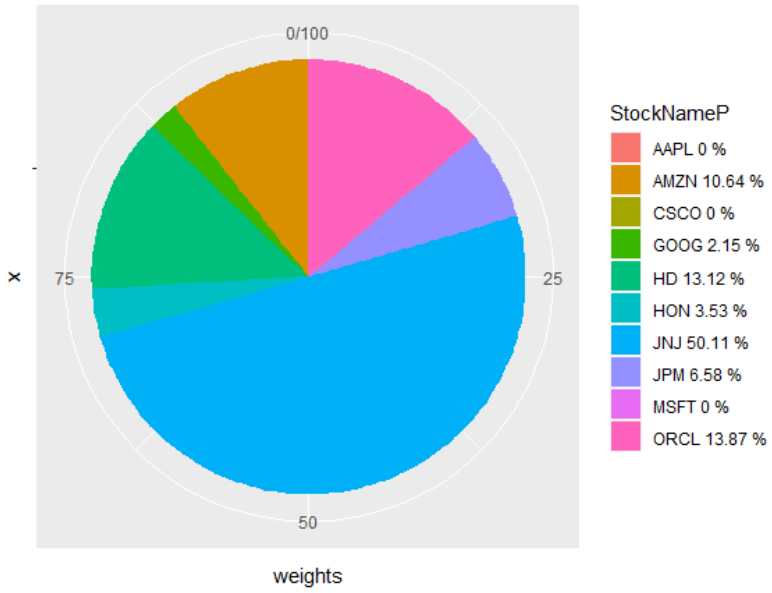


Chart 13: Minimal risk portfolio chart

2. Portfolio based on maximum Sharpe ratio (Blue Triangle):

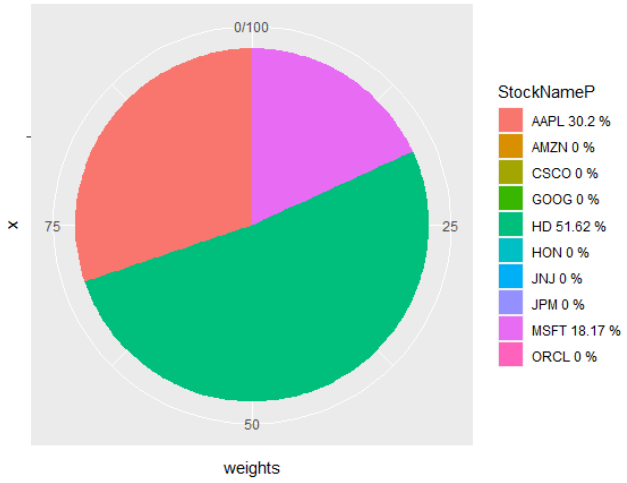


Chart 14: Maximum Sharpe Ratio Portfolio Chart

Conclusion

Different from data analysis, in real securities trading, it is crucial for investors to grasp the timing

of buying/selling (buy/sell signals in quantification). Rather than the error considered in the data analysis. The information obtained by using the prediction model is only for reference. In theory, there is no prediction method with a 100% success rate. Please keep in mind that investment is risky, and you need to be cautious when entering the market.

Based on the previous analysis, the position portfolio under the maximum Sharpe ratio is not the optimal investment ratio during the investment portfolio studied in this article and the historical data period of the study. As investors who pursue the minimization of risks and the maximization of benefits, the first position scheme based on the minimum risk of Markowitz's effective boundary is the optimal choice within the scope of this article. It is inaccurate in this report to use past historical data as the basis for future yield expectations and variance. The future is unpredictable. The most important thing for investors is not to chase returns, but to manage risks and avoid falling into a trap called poverty.

References

- Efficient_frontier https://en.wikipedia.org/wiki/Efficient_frontier
- Residual: https://en.wikipedia.org/wiki/Errors_and_residuals
- How to Check Stationarity of Time Series data in R, 07.31.2021,https://koalatea.io/r-check-stationary/
- ARIMA model Arima Model in R, https://www.educba.com/arima-model-in-r/
- Portfolio Optimization with R/Rmetrics by Diethelm Würtz Yohan Chalabi William Chen Andrew Ellis
- https://en.wikipedia.org/wiki/Markowitz_model
- Code: https://github.com/Troy0207/JiahaoChen-MIS-64060.git