Time Series Analysis in Stock Market and its Application to Portfolio Optimization

by

YIN, Zhendong
SID: 20493538

Department of Mathematics
School of Science

A Project Report Submitted to the
School of Science for the Study Leading to a Project Report of
MATH4999: Independent Capstone Project
in Partial Fulfilment of the requirements for the
Award of the Degree of Bachelor of Science in Mathematics
of The Hong Kong University of Science and Technology

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Supervisor:
Dr. LEUNG, Chi Man

Department of Mathematics
School of Science
chimanleung@ust.hk

May, 2021

Acknowledgement

I'd like to express my great gratitude to my supervisor Dr. LEUNG, Chi Man, whose unlimited steadfast support and inspirations have helped me successfully accomplish this project with excellent outcomes. In a very special way, I thank him for every support he has rendered unto me to help me achieve the goals in this challenging study which brings me a precious research experience.

Special thanks go to my friends and families who have contained the hectic moments and stress I have been through during the course of the research project.

I also thank the school of science and department of mathematics for giving me the grand opportunity to work individually on a particular area that I'm interested in, which get myself prepared for my future research and career.

Contents

Chapter 1
Introduction

1.1 Background

The financial market, with its high uncertainty and rapid changes, often makes a large number of traders and investors confused or even panic. While at the same time, the market composed of traders and traded objects generates massive amounts of data every trading day. Take the Nasdaq index as an example. In each trading day, tens of millions of traders would trade on thousands of stocks, generating trillions of trading volume.

How to dig out the truly valuable part from such massive amount of data? What kind of research methods should be used to process the data, so as to make more reliable predictions on its future trends? How to optimize the investment portfolio so that it can achieve more stable returns and stand out in the market? These problems are widely discussed and analyzed by the researchers and scholars, who are trying to figure out by fitting certain quantitative models as well as implementing advanced analytical methods.

Among all the mathematical models, time series model turns out to be a reliable model. Time series analysis was originally introduced early in 1920s and 1930s, and the systematic work was done by G. E. P. Box and G. M. Jenkins in 1970. The further study has also been developed for several decades. The models and ideas contained in time series analysis make it a classic model that has been widely studied and applied. It has the advantages of analyzing the historical performances of the stock price or return and making further forecast, so as to provide investors with the key investing index to construct portfolio dynamically based on their trading strategies. Time series analysis is not only applied in stock market, due to its endeavor of extracting valuable summary and statistical information from points arranged in chronological order, it's also widely used in collecting, recording and visualizing data in the field of medicine, weather, economics, etc.

1.2 Main Objective

The main objective and goal of this capstone project is to provide the readers with the fundamental logic to construct a bottom-up dynamic portfolio which consists of 4 technological stocks carefully selected from the US and Hong Kong market. While the time series analysis would be used to analyze these stocks in order to back up the design of the portfolio.

Furthermore, in this report, I would also paint out the detailed and comprehensive image of implementing the time series analysis on the historical data of various stocks. The analysis would start from the data transformation and processing, then progress to the time series model fitting, while in this project, the advanced model, ARMA-GARCH model would be adopted. And finally, use the models that have been fitted to each stock to give interpretation and forecasts on their future performances, or more mathematically, the prediction on the return μ and volatility σ of the stock, within one trading day.

1.3 Scope of the study

Since the human society entered the 21st century, the global stock market has experienced two major crises in the past few decades. One was the financial crisis in 2007-2008 which was caused by the bankruptcy of subprime mortgage in the United States, and the other was the global economic recession and stock market circuit breaker triggered by the covid-19 epidemic in 2020. As it has been stated previously, time series analysis would focus on the historical data, especially the informative parts, in order to give more accurate interpretations on the future trends. Since the basic idea of the project is trying to summarize the general patterns and characters of a particular stock as well as making forecasts by fitting time series models to them. As a matter of fact, our study scope would mainly focus on the years between 2009 and 2019 with the time horizon of 11 years.

By wiping out the extreme stock market data generated by the two financial crises, the data for the eleven years from 2009 to 2019 can better reflect the performance of stocks under normal and ordinary economic conditions, while in this case, there would be fewer extreme cases which could interfere sampling and model fitting, thus providing us with data samples with higher reference value.

1.4 Significance of the study

This project would conduct a detailed time series analysis on the real trading data of the stocks to explain the bottom-up process of how to fit the models and apply the models to make forecasts, which could be useful for learners or readers to better understand and learn the logic and ideas behind. What's more, the application of the time series analysis on the portfolio optimization that has been proposed in the project could be incorporated by the traders and investors in their trading strategies to construct dynamic portfolios with better performance. Besides, since the time series models that has been constructed in this project are derived by analyzing the data in the most recent past decade, which could directly benefit on the generalization of features of the stocks in a more quantitative way.

Chapter 2
Methodology

2.1 Data Collecting and Processing

This study is conducted mainly based on the quantitative research on the historical data of different stocks, including those on Hang Seng Index as well as those on Nasdaq in US market. Just as it has been stated in 1.3 Scope of the study, the trading data that has been collected is ranging from Jan 1st, 2009 to Dec 31st, 2019, with approximately 2540 days of trading for each stock. The datasets are download from yahoo.com with the R packages of Quandl, which is an API Wrapper for Quandl.com, together with the R function of stockDataDownload() in the package portfolioBacktest.

The dataset downloaded from the internet using the API wrapper is consisted of daily open price, daily close price, daily trading volume, highest and lowest price of the day. In this project, since the purpose is to analyze the historical performances of the stocks, the daily close price is used for transformation and setting up time series models, and the detailed processing logic and procedure would be introduced and presented in Chapter 3. Meanwhile, the trading volume for a particular stock is a crucial index for reference in order to judge whether the stock is being actively traded, in another word, if the trading volume of a stock reaches a certain level, we believe that interference factors such as artificial market manipulation in this stock can be ignores. Otherwise, the historical price fluctuations of the stock cannot reflect its the normal performance.

2.2 Method of analyzing

After the close prices of the stocks are collected, they are transformed into daily log return for time series modelling. And then, the time series analysis together with the portfolio analysis are conducted with the combination method of mathematical modelling and algorithm programming.

The programming language that is used for model fitting, statistical testing, portfolio analyzing is R, while the statistical software we've adopted is RStudio.

The statistical tests performed in this project includes Unit-Root test, Ljung-Box test, t-test, unit root test, etc.

The graphs and the images are plotted using RStudio.

Chapter 3
System Study, Analysis and Design

3.1 Features of Stock Data

Stocks, like all the financial instruments, the prices are partly determined by the deals enclosed in the market during the trading time while the fundamental logic would be the demand and supply. But in this project, instead of discovering the micro relationships between the price and trading, we would regard the historical performances of the stock as time series.

Generally, there would be upper trend or downside in the stock data, while the trend doesn't have to be linear. Because of this pattern, some stochastic processes were used to generalize the trend, supposing that the performance of the stock could be randomly distributed. The theories of this kind were originated from the earliest study of stock prices. According to the research papers of Regnault (1863) and Bachelier (1900), they raised that price changes D_i could be treated as independent random variables. The probability of price rise and fall is the same. If the Brownian motion of continuous time is known in a uniform time interval, a random walk with a normal increment $N(0, \sigma^2)$ will be obtained. However, in the long-term analysis, there would be much more factors and indicators that would trigger the dynamic fluctuations of the stock price, such as the political, economic or historical events. The time series analysis emphasizes on the correlation between the past and present since there would be lags between the occurrence of the event and the changes caused by it, which could be comprehended as causality factors. And then time series analysis would try to generalize the correlation and lags in terms of statistical methods.

There's also seasonal factor in some certain time series, which means that the movements and changes of the stock follow a fixed and known frequencies. The seasonality could be affected by the seasonal market demand changes and this character often occurs on some business or industry that mainly focus on some particular time. For example, the time series of the sales of ice-cream. The increasing trends always happen in summer and the number would gradually decrease during autumn and winter. In most stocks, the influence of seasonal factor is relatively small, since the price of the stock is partly determined by the market trading conditions and the behaviors of the investor and traders would not be much affected by seasonal changing.

Cyclic is slightly different from seasonality. The stock market will have daily cyclical characteristics, that is, the trading volume at the beginning and end of each trading day will tend to be large, while the trading volume at noon will be small. This cyclical feature is also called U-shaped feature by stock exchanges. But in our time series, the research object would be the prices, close price in particular, of the stocks instead of the trading condition of that day.

3.2 Transformation on the Stock Data

3.2.1 Stationarity

A stationary time series (strictly stationary) $\{x_1, x_2, x_3, \dots \dots\}$ is defined that the mean Ex_t and the joint distribution $(x_{t+h_1}, \dots \dots, x_{t+h_m})$ does not depend on $t \geq 1$ for every $m \geq 1$ and $0 \leq h_1 < \cdots < h_m$, in another word, those fixed patterns and values exist in a stationary time series. Also in a stationary time series, the $Cov(x_t, x_{t+h})$ at lag h is also fixed and doesn't depend on t and is calculated by $\rho_h = \frac{\gamma_h}{\gamma_0}, h \geq 0,$ which is the ***autocorrelation function***, ACF, of the time series $\{x_t\}$.

To illustrate the definition of stationarity, Figure 3.1 plots two different sets of time series. The upper one is the time series of the price of the Tencent stock, and the other is the log return time series.

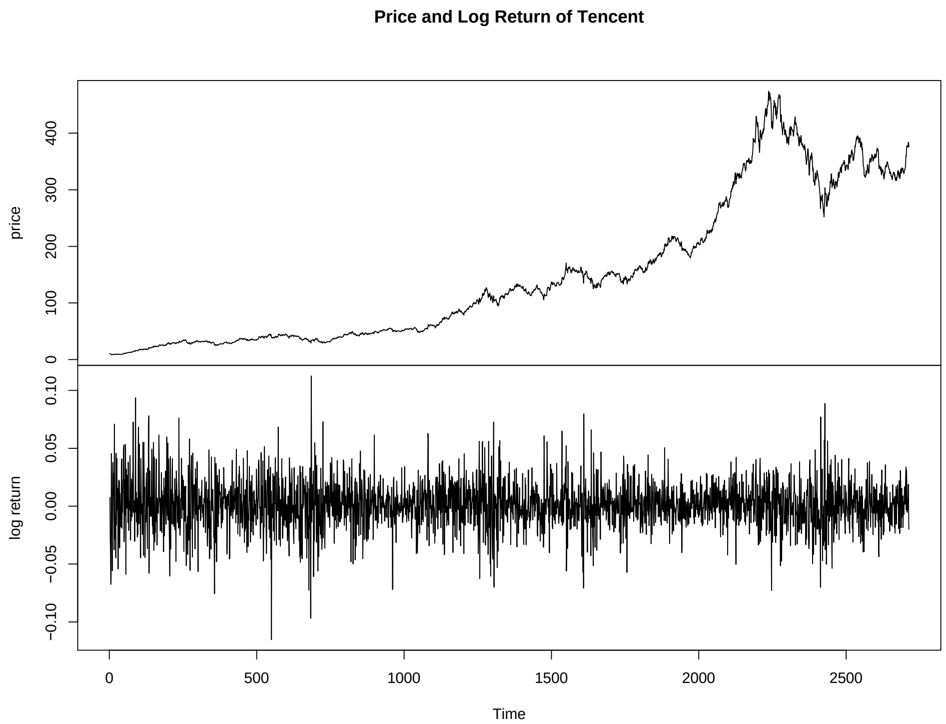


Figure 3.1: Two time series plots of the price and log return of Tencent stock

The price of the stock apparently has its upper trend in the first 2000 days of data, which means the Ex_t of the price is not fixed and gradually increasing, so we could easily distinguish that it's not a stationary time series. The log return of the time series seems to be much more stationary than the price of stock, since the values are mostly concentrated on some constant level close to 0 and has no obvious trend. However, we couldn't easily observe the covariance of these two different time series, so we have to use some statistical tests and the detailed transformation method would be introduced in the upcoming sections.

3.2.2 Difference and Logarithm Transformation

In general, there are two types of return on the stock. The ordinary percentage return $r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100\%$, where P_t , P_{t-1} are the price of the stock of two consecutive trading days. The difference transformation on the original price of the stock usually can help stabilize the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) the trend and seasonality.

Another transformation, logarithm transformation, is widely used in the stock data process. Since the distribution of the ordinary return could be heavy-tailed and the variance is relatively large. With the transformation term, $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$, the variance of the time series could be stabilized and be concentrated to μ. In this case, the log return time series could be more stationary than the ordinary return.

The reason for imposing these transformations and analyze their differences is that a stationary series is relatively easy to predict, since we could simply predict and deduce that its statistical properties will be the same in the future as they have been in the past. But for a non-stationary time series, the fluctuation and volatility are so high that its historical properties couldn't be easily reflected and measured using statistical models. In order to have more accurate forecast on the performance of the stocks, these transformations are extremely useful and necessary.

Example: Time series of prices and log returns of nine stocks

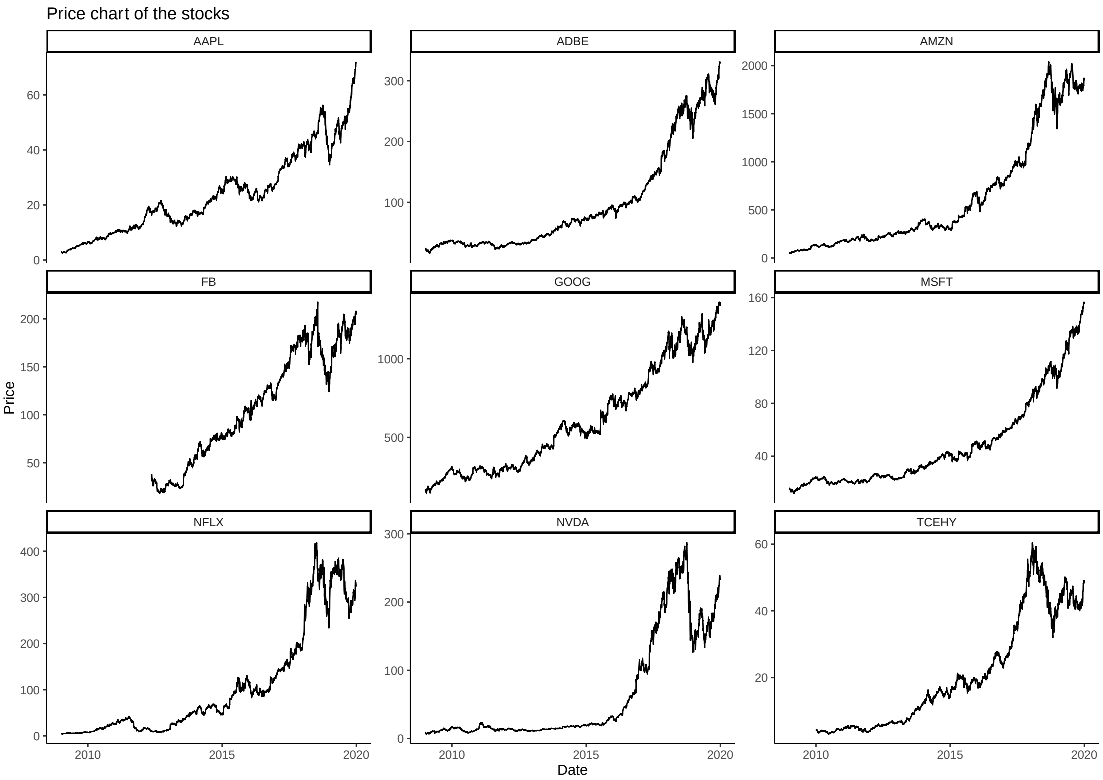


Figure 3.2: Plots of the price of 9 stocks

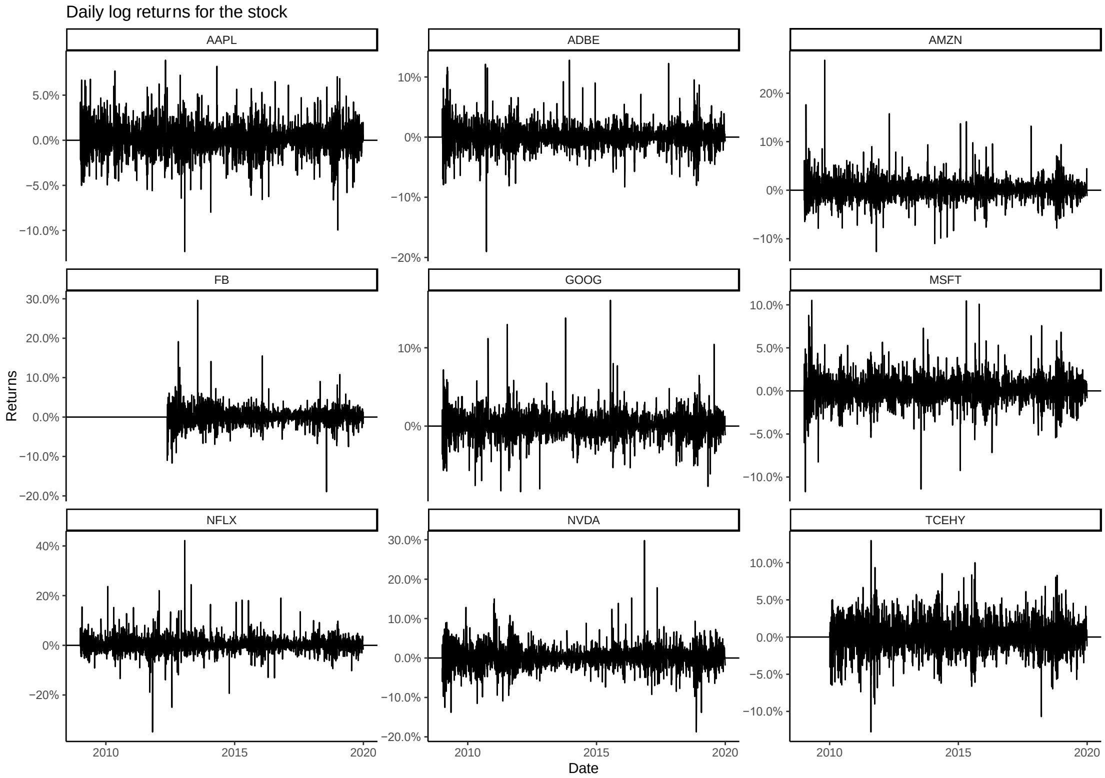


Figure 3.3: Plots of the log return of 9 stocks

The comparison could be shown between Figure 3.2 and Figure 3.3. During 2009 and 2019, those stocks generally have upper trends, but also have high fluctuations. After imposing the logarithm transformation, although some volatile patterns could still be observed, which are called "volatility clustering" and the detailed explanations about it would be given in later sections, the time series are more stationary, therefore, the fitted model as well as the predictions would be more accurate and convincing.

3.2.3 Unit Root test

Continuing on the last section, I'm introducing the statistical test, unit root test, which is used to test the stationarity of a time series. Generally, a particular time series consists of 3 parts, the first is the deterministic component, which could be the trend and seasonal components of the time series and we usually fit this part with certain time series models. The second part would be the stochastic component with random process. The last part is a stationary error time series. The goal of the unit root test is to check whether there is any unit root in the stochastic component of a time series. If a unit root does exist, then it shows a systematic character that is unpredictable, so that we could no longer fit any linear models to make forecast on the trend. And the conclusion is the time series is non-stationary with presence of unit root.

Augmented Dickey-Fuller test

For a time series $\{x_t\}$, let B denote the backshift operator which is define by $Bx_t = x_{t-1}$, $B^2x_t = Bx_{t-1} = x_{t-2}$,..., $B^m x_t = x_{t-m}$. Based on the time series that we would introduce in the next sections, consider AR(p) model, $\phi(B)x_t = \mu + \epsilon_t$, where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$. Define $\Delta x_t = x_t - x_{t-1}$, the AR($p$) model can be written as

$$x_t = \mu + \beta_1 x_{t-1} - \sum_{j=1}^{p-1} \beta_{j+1} \Delta x_{t-j} + \epsilon_t , \qquad \text{where } \beta_j = \sum_{i=j}^{p} \phi_i \tag{3.1}$$

$\phi(1) = 1 - \phi_1 - \phi_2 - \cdots - \phi_p = 0$, if and only if $\beta_1 = \sum_{i=1}^{p} \phi_i = 1$, so the unit root test has the null hypothesis H_0: $\beta_1 = \sum_{i=1}^{p} \phi_i = 1$. The *Augmented Dickey-Fuller test* has the test statistics

$$ADF = \frac{\widehat{\beta_1} - 1}{\widehat{se}(\widehat{\beta_1})} \tag{3.2}$$

$\widehat{\beta_1}$ is the OLS estimate of β_1 in the regression model (3.1), while $\widehat{se}(\widehat{\beta_1})$ is the estimated standard error. If null hypothesis holds, then the conclusion of the time series being non-stationary could be drawn.

3.3 Basic Time Series models

3.3.1 ACF and PACF

The autocorrelation function, ACF, which has been introduced in 3.2.1, for a time series $\{y_t\}$, the ACF would be written as

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=k+1}^{T} (y_t - \bar{y})^2} \tag{3.3}$$

In general, ACF describes the degee of correlation between the current value of the series and its past value. Time series can contain trends, seasonality, periodicity and residuals. ACF considers all these components when looking for correlations, so it is a "complete autocorrelation graph".

Partial correlation is conditional correlation, which focus on interpretating the correlation between two lags and also take the other lags into consideration. To illustrate that, the 1st order PACF equals to the 1st order correlation. The 2nd and 3rd order PACF is written as

$$\frac{Cov(x_t, x_{t-2} | x_{t-1})}{\sqrt{Var(x_t | x_{t-1}) \cdot Var(x_t | x_{t-2})}} \tag{3.4}$$

$$\frac{Cov(x_t, x_{t-3} | x_{t-1}, x_{t-2})}{\sqrt{Var(x_t | x_{t-1}, x_{t-2}) \cdot Var(x_{t-3} | x_{t-1}, x_{t-2})}} \tag{3.5}$$

And so on, for any lag and any order of the time series. In the upcoming sections, the auto-regression model, AR model, and moving average model, MA model, will be introduced. For an AR(p) model, the plot of the PACF would cut off at lag p, and the previous lags could be regarded as significant. And for an MA(q) model, the ACF usually has a cut off at lag q. The ACF and PACF could be calculated and plotted with the R function of acf() and pacf(), plot().

Figure 3.4 shows the ACF and PACF of the log return of Tencent, the plots are essential for model construction when determining the number of lags. And the detailed process of lags choosing and model construction would be explained in the upcoming example of Practical model fitting on Tencent.

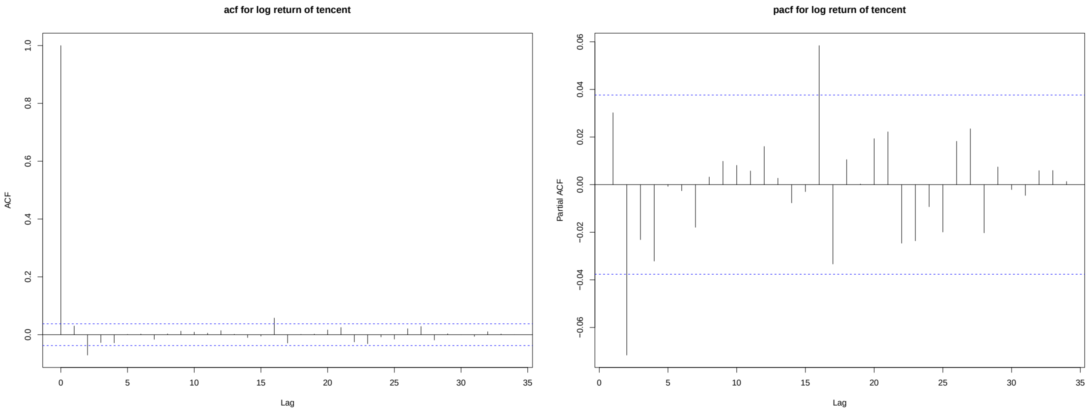


Figure 3.4: ACF and PACF of the log return of Tencent stock

3.3.2 Test of independence: Ljung-Box test

In section 3.2.1, one of the characters of time series $\{x_t\}$, stationarity, has been analyzed. The independence inside $\{x_t\}$ is another main feature. Specifically, the data of $\{x_t\}$ are independently distributed, that is, the correlations form the sample are 0, thus the observed correlations in the data result from randomness sampling process. $\hat{\rho}_h$ is asymptotically independent and have the distribution $N(0,\frac{1}{n})$ as $n \to \infty$. The null hypothesis $H_0: \rho_0 = \cdots = \rho_m = 0$ would be rejected if $|\hat{\rho}_h| \geq z_{1-\alpha}/\sqrt{n}$. Moreover, Ljung-Box test is widely adopted in this independent testing process using the test statistic $Q(m)$, where

$$Q(m) = n(n+2)\sum_{h=1}^{m} \hat{\rho}_h^2 /(n-h) \tag{3.6}$$

n is the sample size and h is the number of lags being tested. Under the null hypothesis, $Q(m) \sim \chi_h^2$. With the significant level of α, the critical region for the rejection of null hypothesis is $Q(m) > \chi_{1-\alpha}^2$. The test could be achieved using R function Box.test with type “Ljung”.

3.3.3 AR and MA model

In the time series, the historical performances are usually taken into consideration so as to make forecasts for the future. The autoregression model, AR model, is one of the time series models that estimates the one step ahead with a linear combination of the past lags. For a time series $\{x_t\}$, the AR(p) could be written as

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t \tag{3.7}$$

ε_t is known as the white noise, which is the type of time series that is totally random and not able to be predicted. The changes of $\phi_1, \phi_2 \dots \phi_p$ will have direct impact on the pattern of the time series, and the coefficients, in another word, the weight on different lags are dynamic because of the refreshment of the daily performance of the stock in this report. But for a particular stock, the lags in the time series model are fixed since the selection of lags is based on the comprehensive analysis on the massive historical data, while those lags are most significant.

Different from autoregression model, the moving average model, MA model, doesn't just simply uses the past value of the time series $\{x_t\}$, instead, it actually makes adjustment on the past forecast with the errors being created. MA(q) is written as

$$x_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{3.8}$$

The values of $\{\varepsilon_t\}$ are not observed but are derived from the former error between the prediction value and the real value.

The AR and MA model can transform with each. To illustrate that, for an AR(1) model, with the form from 3.7

$$\begin{aligned} x_t = \phi_1 x_{t-1} + \varepsilon_t = \phi_1(\phi_1 x_{t-2} + \varepsilon_t) + \varepsilon_t &= {\phi_1}^2 x_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= {\phi_1}^3 x_{t-3} + {\phi_1}^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \varepsilon_t + \phi_1 \varepsilon_{t-1} + {\phi_1}^2 \varepsilon_{t-2} + {\phi_1}^3 \varepsilon_{t-3} + \cdots \end{aligned} \tag{3.9}$$

The newly transformed model 3.9 could be regarded as MA(∞) model. The fundamental logic in this transformation is that in AR(1) model, one historical lag is used to estimate the performance. But the lag can be further decomposed into previous lags and the historical estimation error. With $0 < \phi < 1$, $\lim_{n \to \infty} {\phi_1}^n x_{t-n} = 0$, the lag term could be eliminated and the model can be constructed with only the error terms.

3.3.4 ARMA and ARIMA model

The ARMA model can be easily comprehended as the combination of the AR and MA model, with the corresponding variables ARMA(p,q) and can be expressed as

$$x_t = c + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{3.10}$$

Since the linear combination is applied on the AR and MA model, the ARMA model would have the same stationarity and transformation conditions stated in the previous section. The advantage of ARMA over AR and MA is that it not only adopts historical data to make forecast but it also makes opportune adjustments on the estimate to try to minimize the error. In this case, it would get a relatively more accurate prediction on the future trend.

The ARIMA(p,d,q) is almost the same as ARMA(p,q), except from implementing an addition of difference transformation on the original data. And the rest of the processes, such as model fitting and residual analysis, are the same as ARMA model. As it has been analyzed in section 3.2.1, the difference method is to stabilize the trend of the original data and make it more stationary, while the logarithm method has the effect of centralizing the data. In this project, the logarithm transformation is adopted and no more difference process is needed.

3.3.5 Forecasting and Order Selection

In order to identify and select the model orders p and q, the information criterion is adopted. Akaike's Information Criterion (AIC) and Bayesian Information Criterion are helpful in the predictor selection for regression and are also useful in the order determination of an ARMA model. We first let $\boldsymbol{\theta} = (c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)^T$. The log-likelihood function is given by

$$l(\boldsymbol{\theta}, \sigma) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t - c - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p}, \dots, -\theta_1\varepsilon_{t-1} - \dots - \theta_q\varepsilon_{t-q})^2 \quad (3.11)$$

And the forms of AIC and BIC in ARMA(p,q) order selection are written as

$$AIC(d) = -2l(\widehat{\boldsymbol{\theta}}, \hat{\sigma}) + 2d$$
$$BIC(d) = -2l(\widehat{\boldsymbol{\theta}}, \hat{\sigma}) + d\ln(n) \quad (3.12)$$

n is the sample size, $d = p + q + 1$, $l(\boldsymbol{\theta}, \sigma)$ is the log likelihood function of the ARMA(p,q) model, $(\widehat{\boldsymbol{\theta}}, \hat{\sigma})$ is the *maximum likelihood estimator*, MLE, of $(\boldsymbol{\theta}, \sigma)$. In order to select the best model, the selection criteria is to minimize the AIC and BIC value of the fitted model and the p and q order could be identified.

After going through data transformation and stationarity testing, the processed data would be used to fit appropriate time series models and make forecasts. With the sample data $\{x_1, x_2, \dots, x_t\}$ regarded as the current and past observations, we assume the initialization

$$x_0 = \cdots = x_{t-p+1} = 0 = \varepsilon_0 = \cdots = \varepsilon_{t-p+1} \quad (3.13)$$

Based on the ARMA(p,q) model, the estimated one-step-ahead forecast $\hat{x}_{t+1|t}$ has the following formula

$$\hat{x}_{t+1|t} = c + \sum_{i=1}^{p}\phi_i x_{t-i+1} + \sum_{j=1}^{q}\theta_j\varepsilon_{t-j+1} \quad (3.14)$$

The h-step-ahead forecast, $\hat{x}_{t+h|t}$, can also be derived by rewriting 3.14, however, some variables would be missed since the forecasts value $\hat{x}_{t+h-1|t}, \dots, \hat{x}_{t+1|t}$ are used to construct the new model, with the coefficient being modified based on the 3.14.

Example: Practical model fitting on Tencent

In the previous contents, the initial image of the whole process of time series analysis has been drawn. In order to illustrate more on it, we choose Tencent stock as the sample data and fit the coherent ARMA(p,q) model. After implementing the logarithm transformation on the stock price, we shall use the ACF and PACF plots of the log return to initially determine the order of the model. As Figure 3.4 has shown, ACF and PACF have cutoffs in lag 1 and 5, so our initial order is chosen as ARMA(5,1), then by changing the variables p and q, the other models ARMA(4,1) ARMA(5,1) ARMA(5,2) ARMA(4,2) are also constructed. By comparing the values of AIC and BIC in **Table 3.1**, it turns out that ARMA(5,2) has relatively better regression outcomes, and the coefficient is given, values in () are the standard error of each estimation. In R, the auto.arima() function could be used to fit and test out the best model, the algorithm of the function would be explained in Appendix B.

$$\begin{array}{c} x_t = 0.0013 - 0.7247x_{t-1} + 0.0872\,x_{t-2} - 0.0828\,x_{t-3} - 0.0378x_{t-4} \\ (0.0004) \quad (2.0209) \qquad (1.8176) \qquad (0.0900) \qquad (0.2069) \\ -0.0328x_{t-5} + \varepsilon_t + 0.7547\varepsilon_{t-1} - 0.1383\varepsilon_{t-2} \\ (0.0719) \qquad\qquad (2.0224) \qquad (1.8794) \end{array}$$

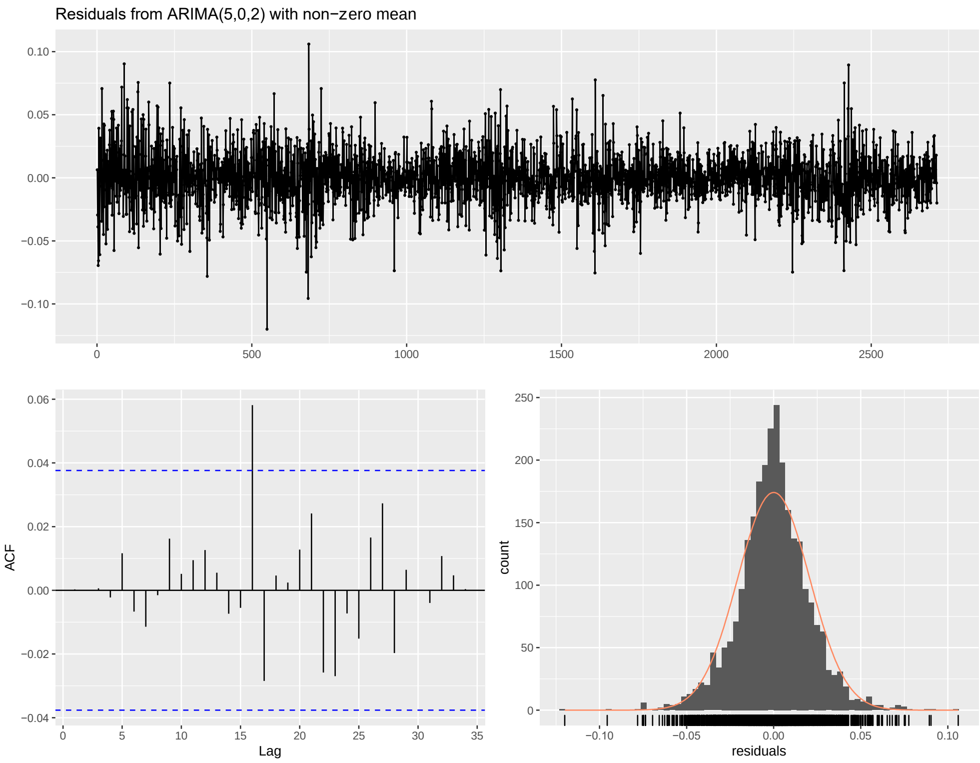


Figure 3.5: Features of residuals of ARMA(5,2) fitting on log return of Tencent stock

Table 3.1. AICs given by ARMA(p,q) (p=4,5; q=1,2).

$p\backslash q$	$q = 1$	$q = 2$
$p = 4$	-13333.39	-13334.88
$p = 5$	-13335.39	-13336.93

After fitting ARMA(5,2) to the data, we derive the residual terms or error terms $\{\varepsilon_t\}$ between the estimated value $\{\hat{x}_t\}$ and the real value $\{x_t\}$. In Figure 3.5, the upper image plots the residuals time series. It also plots the ACF of the residuals, from which it could be clearly observed that the correlations of the residuals in the first 15 lags are much smaller than the original time series, which means the residual terms are independent and stationary. To check that statistically, recall section 3.3.2, the Ljung-Box test is used to check the whether the residual terms are white noise.

The test statistic value of the residuals is $Q(m) = 1.9032$, with the $p\ value = 0.5927 > 0.05$, so the null hypothesis is accepted with $\rho_0 = \cdots = \rho_m = 0$. So the residual terms could be regarded as white noise which means the model fitting is successful.

The one-step point forecast value $\hat{x}_{t+1|t} = 0.00052584$, with 95% confidence interval $[-0.04000060\ 0.04105228]$. In order to make more comprehensive and accurate forecasts, more advanced model shall be introduced and constructed.

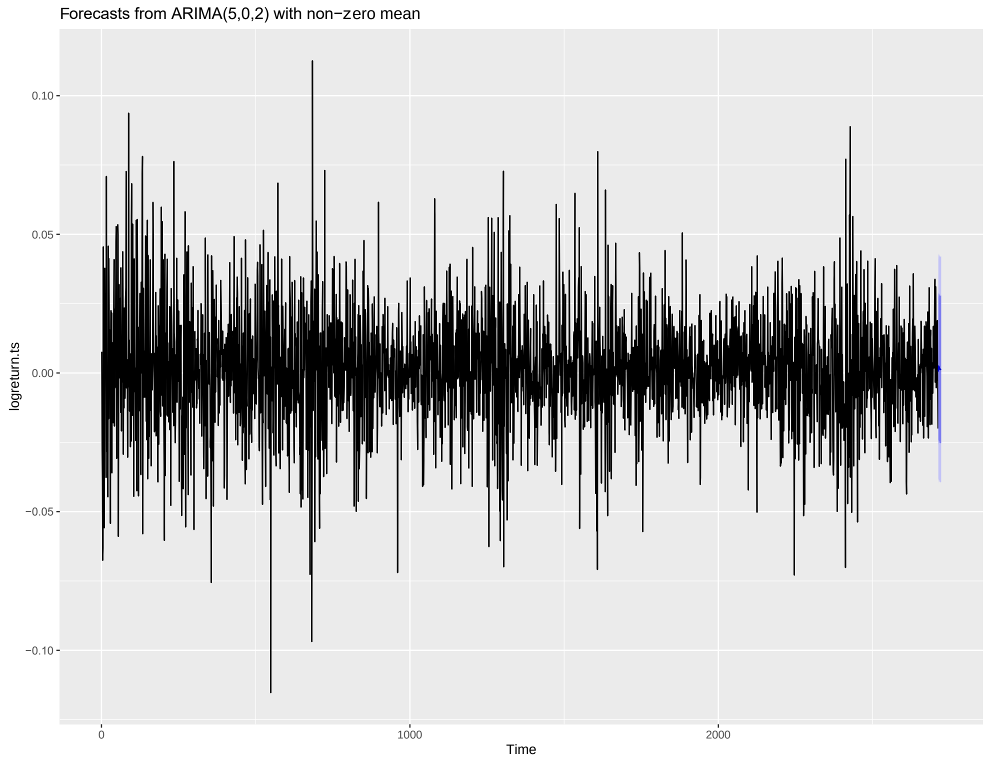


Figure 3.6: Plot of forecasts (blue lines) of ARMA(5,2) fitting on log return of Tencent stock

3.4 Advanced Time Series models

3.4.1 Features of volatility

In the previous models, we mostly focus on the log return modelling and forecasts, however, there actually exists other informative data inside the original time series that is known as the volatility, σ, which can also be modeled and generalized the patterns in order to make better forecasts. In this chapter, the volatility modelling is the main topic.

Start from the patterns of the volatility of the stock. In time series, σ_t denotes the volatility of r_t at time *t*. And the estimate of $\sigma_t{}^2$, is give as

$$\hat{\sigma}_t{}^2 = \frac{1}{k-1}\sum_{i=1}^{k}(r_{t-i} - \bar{r})^2 \qquad (3.15)$$

$\bar{r} = \sum_{i=1}^{k} r_{t-i}/k$, is the mean return within the window size k. The volatility estimate $\hat{\sigma}_t{}^2$ given by (3.15) is called the *k-day historic volatility.*

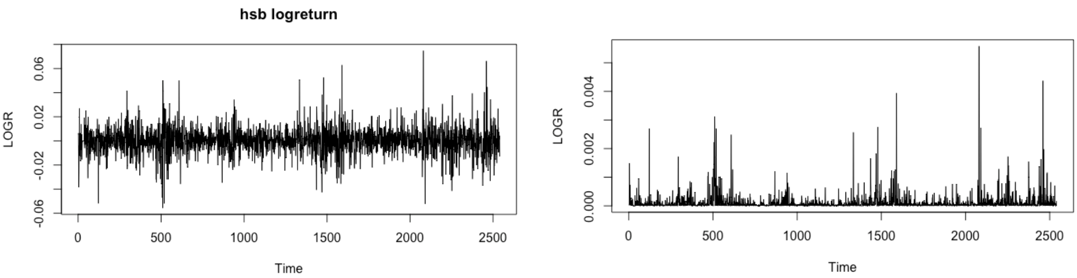


Figure 3.7: Left panel: log return of HSB; Right panel: volatility of HSB with k=5

Notice that in Figure 3.7 left panel, the log return exhibits the feature that some large return/losses are always concentrated to form clusters, this phenomenon is called *volatility clustering*. This phenomenon would result in the fact that the squared log return, which can be partly regarded as volatility since the mean $\bar{r}$ is close to 0 has larger autocorrelations than the original log return, just as what has been shown in the right panel.

3.4.2 ARCH and GARCH model

In this chapter, the objective is to generalize and estimate the current volatility with certain models. Just like in the previous chapter, we adopts AR, MA, ARMA models to estimate the return, the *autoregressive conditional heteroskedastic*, ARCH model with order *k* is used and written as

$$u_t = \sigma_t \epsilon_t, \qquad \sigma_t{}^2 = \omega + \sum_{j=1}^{k} \alpha_j u_{t-j}{}^2 \qquad (3.16)$$

Where $u_t = r_t - \bar{r}$, so ϵ_t, which are i.i.d. random variables with mean 0 and variance 1 and have the standard normal distribution a standardized Student t-distribution, are used to randomize variable u_t. And the later function looks like the general MA model, which use the error terms to regress on the $\sigma_t{}^2$. The practical steps of constructing an ARCH model are stated below:

Step (1): establish a mean equation by testing the serial correlation of the data, and if necessary, establish a time series model (such as ARMA) for the yield series to eliminate any linear dependence.

Step (2): test the ARCH effect on the residual of the mean equation

Step (3): if there is ARCH effect, the volatility model is established

Step (4): verify the fitted model and optimize it if necessary

The ARCH effect can be comprehended as the independency of the volatility. Recall in 3.3.2, Ljung-Box test is used to test the independency of the time series, so the ARCH effect test's objective is the squared log return time series, with null hypothesis $H_0 : \rho_0 = \cdots = \rho_m = 0$. Specifically, ARCH-LM test is the standard test to detect autoregressive conditional heteroscedasticity. (In statistics, a sequence of random variables is homoscedastic if all its random variables have the same finite variance. This is also known as homogeneity of variance. The complementary notion is called heteroscedasticity.)

The *generalized ARCH* model, also known as GARCH model, has the same condition as ARCH, but absorb more lags, which is similar to the ARMA model, and the form is written as follow

$$u_t = \sigma_t \epsilon_t, \qquad \sigma_t{}^2 = \omega + \sum_{i=1}^{h} \beta_i \sigma_{t-i}{}^2 + \sum_{j=1}^{k} \alpha_j u_{t-j}{}^2 \tag{3.17}$$

The same stationarity assumptions of ARMA model also apply to GARCH model. The α β which are nonnegative also has the following constraints to ensure that u_t is covariance stationary

$$\sum_{j=1}^{k} \alpha_j + \sum_{i=1}^{h} \beta_i < 1 \tag{3.18}$$

The most widely used GARCH model in estimating the volatility of financial time series is GARCH(1,1) model $\sigma_t{}^2 = \omega + \alpha\sigma_{t-1}{}^2 + \beta u_{t-j}{}^2$. But in more cases in our project, GARCH(1,1) isn't enough, so more models with higher orders are applied. The practical steps of constructing an GARCH model are stated below:

Step (1): establish a mean equation by testing the serial correlation of the data, and if necessary, establish the best AR model $x_t = c + \sum_{i=1}^{q} \phi_i x_{t-q} + \varepsilon_t$

Step (2): compute and plot ACF of ε_t, the residual terms

Step (3): test the lags significant using Ljung-Box test to fit the appropriate GARCH model candidates by changing the order of p, significant level could be 10%.

The whole process of fitting ARCH and GARCH models could be achieved in R using function Archfit(), and Garchfit().

3.4.3 ARMA-GARCH model

Concluding the time series models we have gone through so far, AR, MA, ARMA models could be used to generalize the trend of the log return time series and furthermore make forecasts on it. ARCH and GARCH models can better trace the movements of the volatility and use statistical method to summarize the pattern and make forecasts on the future volatility.

So in this project, we finally adopt ARMA-GARCH model, which is the combination of ARMA and GARCH model. For the log return time series $\{r_t\}$, ARMA is used to estimate the log return performances of the stock with one-step-forecast $r_{t+1|t}$. While the GARCH model will give us the one-step-forecast $\sigma_{t+1|t}$. Continuing on the practical example ARMA(5,2) that has been fitted to Tencent stock log return, the GARCH model of the volatility is given as

$$\sigma_t^2 = 0.000017 + 0.3713\sigma_{t-1}^2 + 0.4765\sigma_{t-2}^2 + 0.1122u_{t-1}^2$$

With order GARCH(1,1) and $u_t = \sigma_t \epsilon_t$. And with the combined ARMA-GARH model, the forecasting outcomes are $\hat{r}_{t+1|t} = 0.001519$, $\hat{\sigma}_{t+1|t} = 0.01648$. Then we perform Monte-Carlo simulation on the forecasting result, assuming that the log return would follow a standard normal distribution with $N(\hat{r}_{t+1|t}, \hat{\sigma}_{t+1|t}{}^2)$

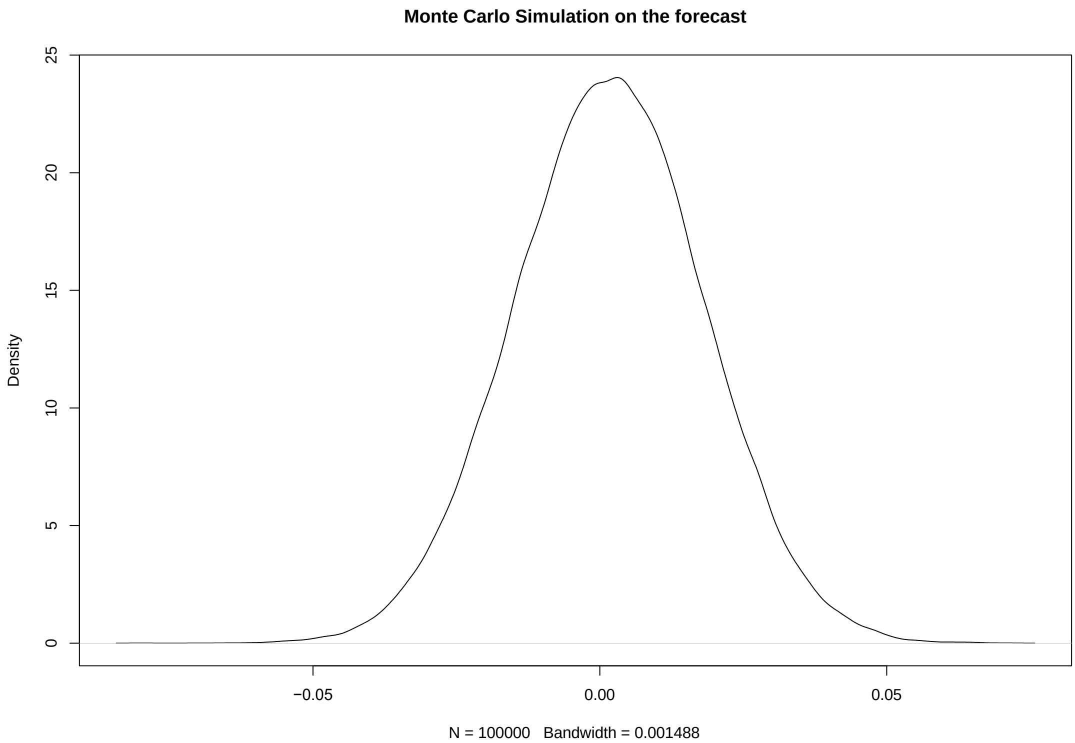


Figure 3.8: Monte Carlo simulation on the forecasting outcome with 100000 runs

3.5 Optimization of the Portfolio

3.5.1 Portfolio types

Portfolio constructing has been developed for decades, since the portfolio shall meet the requests from investors, a lot of criteria have been set, such as, risk, annualized rate of return, Sharpe ratio, maximum drawdown, etc. To satisfy the need of different needs of investment goal, various of portfolios are established with mathematical tools.

Markowitz's mean-variance Portfolio

Markowitz's mean-variance portfolio (MVP), is to find a tradeoff between the expected return $w^T\mu$ and the risk of the portfolio $w^T\Sigma w$, based on the risk-averse value λ of different preferences. MVP is to maximize

$$w^T\mu - \lambda w^T\Sigma w \tag{3.19}$$

$$\text{Subject to } \mathbf{1}^T w = 1, w \geq 0$$

While w is the weight distribution of the portfolio, Σ is the covariance matrix, μ is the return matrix.

Maximum Sharpe ratio Portfolio

Sharpe ratio is defined as $S_a = \frac{E[R_a - R_b]}{\sigma_a}$, R_a and σ_a are the return and risk of the risky asset while R_b is the return on the risk-free asset. Maximum Sharpe ratio portfolio (MSRP) is to maximize

$$\frac{w^T\mu}{\sqrt{w^T\Sigma w}} \tag{3.20}$$

$$\text{Subject to } \mathbf{1}^T w = 1, w \geq 0$$

Risk Parity Portfolio

Risk parity portfolio (RPP) adopts the allocation strategy that uses risk to determine allocations across various components of an investment portfolio so as to achieve the risk parity balance in the portfolio.

The variance of the portfolio $\sigma(w) = \sqrt{w^T\Sigma w}$, which can be expressed as

$$\sigma(w) = \sum_{i=1}^{N} w_i \frac{\partial\sigma}{\partial w_i} = \sum_{i=1}^{N} \frac{w_i(\Sigma w)_i}{\sqrt{w^T\Sigma w}} \tag{3.21}$$

The risk contribution (RC) from (3.21) is

$$RC_i = \frac{w_i(\Sigma w)_i}{\sqrt{w^T\Sigma w}} \tag{3.22}$$

The risk parity portfolio is to achieve that $RC_i = \frac{1}{N}\,\sigma(w) = \frac{\sqrt{w^T\Sigma w}}{N}$. Subject to $\mathbf{1}^T w = 1$, $w \geq 0$. The construction of risk parity portfolio could use **R** package **riskParityPortfolio**.

Global Minimum Variance Portfolio

Global Minimum Variance Portfolio (GMVP) aims to achieve the minimum variance in the portfolio and to minimize $\sigma(w) = \sqrt{w^T \Sigma w}$, Subject to $\mathbf{1}^T w = 1, w \geq 0$. The weight of the GMVP is then written as

$$w_{GMVP} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}} \tag{3.23}$$

3.5.2 Simulation and Back-testing on the Portfolios

Initially, based on the historical data of the 8 stocks ("GOOG", "NFLX", "AAPL", "AMZN", "FB", "NVDA", "TCEHY", "ADBE"), we first conduct some simulations and back-testing to illustrate more on the portfolio construction. The μ and Σ are calculated based on 240 trading days, and the portfolios are optimized and refreshed every 60 trading days.

All the portfolio strategies that are introduced in 3.5.1 are tested, while in Markowitz's mean-variance portfolio (MVP), λ takes 3 different values of 0.25, 0.5,0.75.

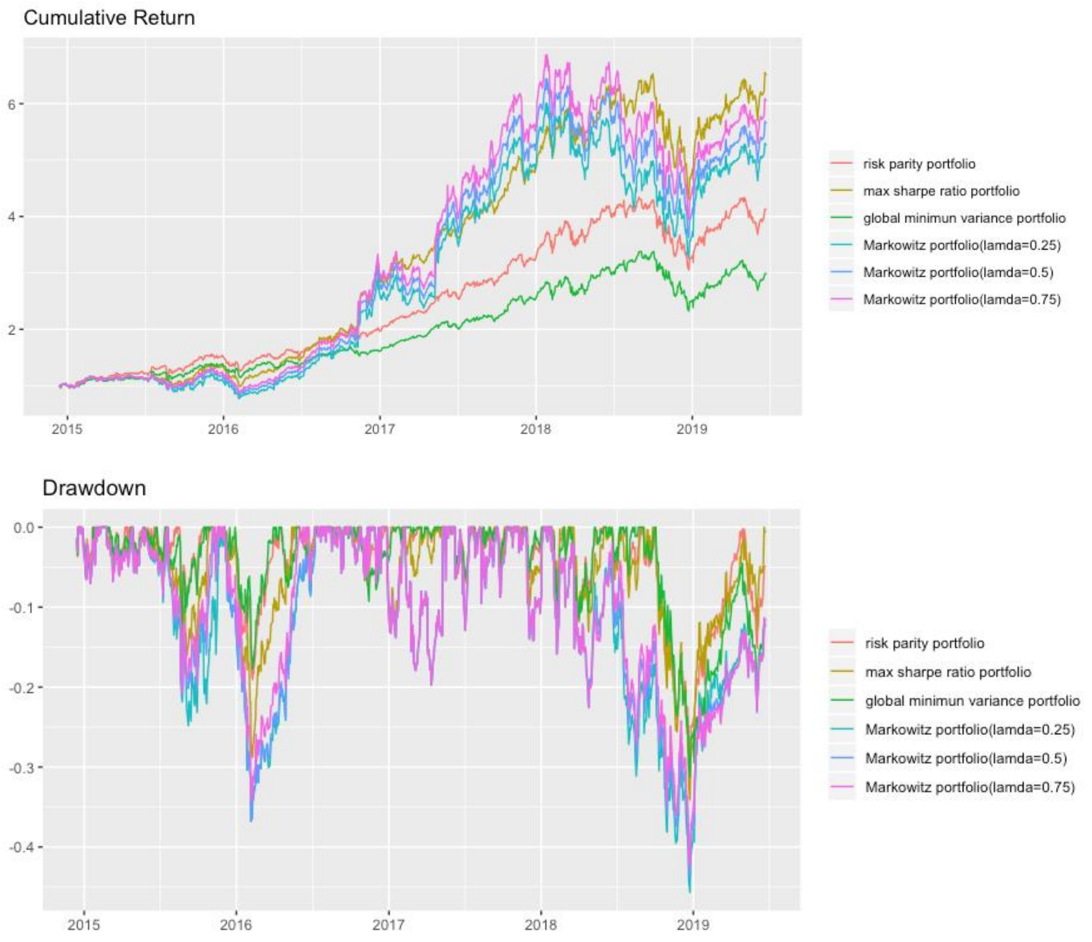


Figure 3.9: Cumulative return and drawdown based on the back-testing of 6 different portfolios

This initial attempt to simulate and back-testing the portfolio strategies provide us a comprehensive image to deeply understand the meaning of these portfolio. While at the same time, by comparing different portfolios' performances, one would get to know the pros and cons of each portfolio strategy. It turns out that for technological stock, the performances the price of the stock could be highly volatile, which bring high risk to these stocks. However, trying to escape from risk won't help much on the investment outcomes. Among all 6 portfolios, the Maximum Sharpe ratio Portfolio (MSRP) has the annualized return of 51.5% with annualized variance of 30.5%, while Global Minimum Variance Portfolio (GMVP) only has the annualized return of 27.4% with the annualized variance of 21.4%.

Although different investors may have different investing preferences, we could still derive the conclusion that for those 8 technological stocks listed on Nasdaq, the Maximum Sharpe ratio Portfolio (MSRP) could present better investment outcomes in the long term. This conclusion could be useful for investors who wants to achieve outstanding return and could also take risks.

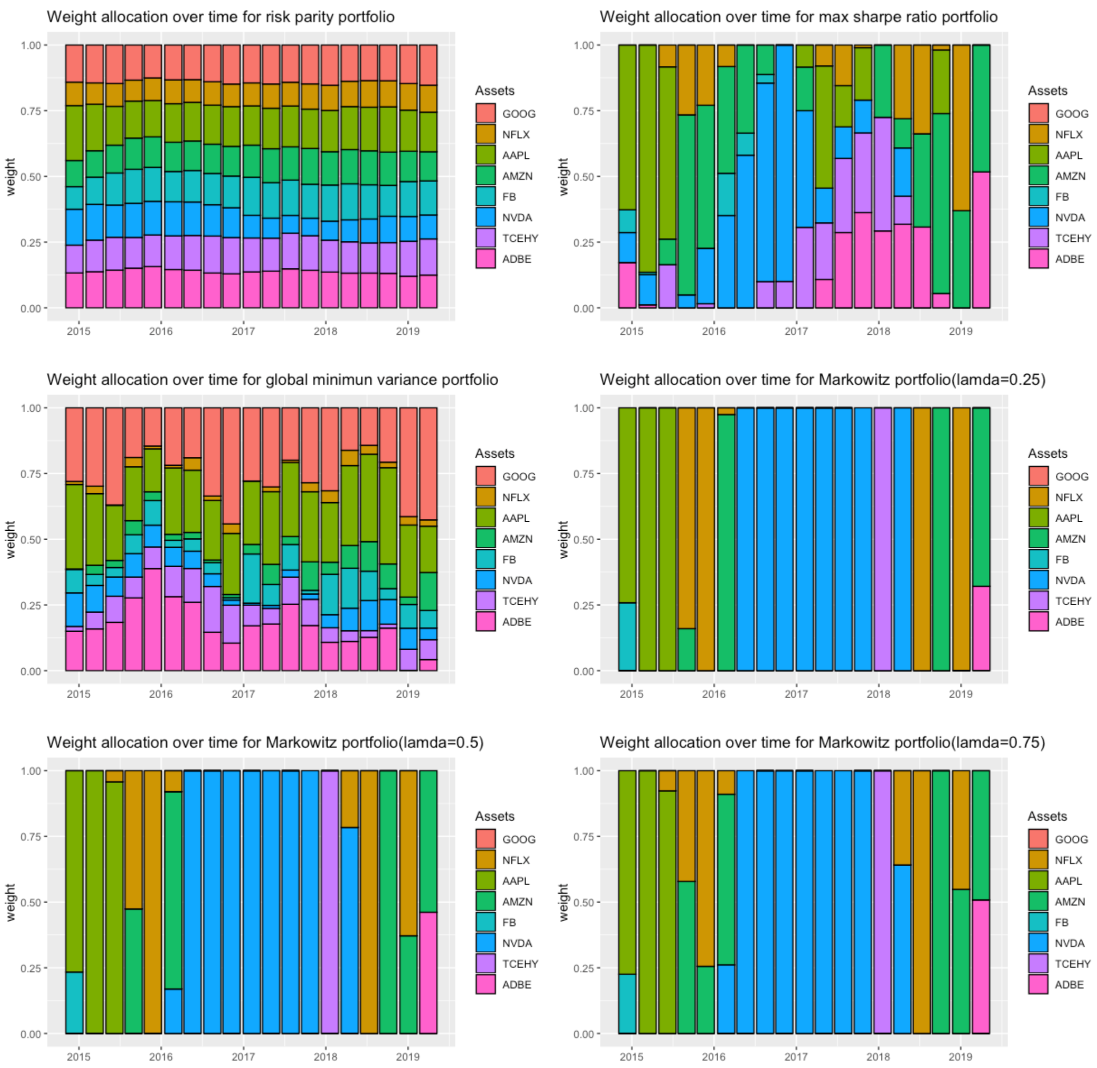


Figure 3.10: Dynamic weight allocation based on the back-testing of 6 different portfolios

Chapter 4
Presentation of Results

4.1Return and Covariance matrix based on ARMA-GARCH

Based on what has been analyze in Chapter 3, we would try to apply the forecasting outcomes based on ARMA-GARCH models of 4 stocks, to different portfolio strategies. Instead of using the historical volatility and return, the ARMA-GARCH would provide a better interpretation of the current condition, thus provide us with more accurate forecast.

First, we have the **Σ** covariance matrix based on the historic performances. The 4 stocks includes google, apple, amazon and tencent, and the **Σ** is given as

$$\mathbf{\Sigma} = \begin{pmatrix} 0.000217 & 0.000115 & 0.000171 & 0.000072 \\ 0.000115 & 0.000612 & 0.000132 & 0.000250 \\ 0.000171 & 0.000132 & 0.000354 & 0.000006 \\ 0.000072 & 0.000250 & 0.000006 & 0.000350 \end{pmatrix}$$

And the fitted ARMA-GARCH models of the 4 stocks are written as

$$\boldsymbol{Google}: \boldsymbol{ARMA(2,2) - GARCH(1,1)}:$$

$$x_t = 0.00077 - 1.2046x_{t-1} - 0.9948x_{t-2} + 1.2088\varepsilon_{t-1} - 0.9982\varepsilon_{t-2} + \varepsilon_t$$

$$\sigma_t^2 = 0.000023 + 0.8230\sigma_{t-1}^2 + 0.0857u_{t-1}^2$$

$$\boldsymbol{Apple}: \boldsymbol{ARMA(2,2) - GARCH(1,1)}:$$

$$x_t = 0.001772 - 1.5185x_{t-1} - 0.8560x_{t-2} + 1.5250\varepsilon_{t-1} + 0.8823\varepsilon_{t-2} + \varepsilon_t$$

$$\sigma_t^2 = 0.000018 + 0.8285\sigma_{t-1}^2 + 0.1098u_{t-1}^2$$

$$\boldsymbol{Amazon}: \boldsymbol{ARMA(2,0) - GARCH(1,1)}:$$

$$x_t = 0.001761 + 0.0302x_{t-1} - 0.0056x_{t-2} + \varepsilon_t$$

$$\sigma_t^2 = 0.00018 + 0.8349\sigma_{t-1}^2 + 0.1045u_{t-1}^2$$

$$\boldsymbol{Tencent}: \boldsymbol{ARMA(5,2) - GARCH(1,1)}:$$

$$x_t = 0.001563 - 0.4122x_{t-1} - 0.9216x_{t-2} - 0.0203x_{t-3} - 0.0799x_{t-4} - 0.0158x_{t-5} + 0.4525\varepsilon_{t-1} + 0.8800\varepsilon_{t-2} + \varepsilon_t$$

$$\sigma_t^2 = 0.000011 + 0.8982\sigma_{t-1}^2 + 0.0757u_{t-1}^2$$

Table 4.1. Forecasted return and variance based on ARMA-GARCH

$stock\backslash forecast$	$\hat{r}_{t+1\|t}$	$\hat{\sigma}_{t+1\|t}$
$google$	0.001530	0.01241
$apple$	0.001541	0.01211
$amazon$	0.001677	0.01224
$tencent$	0.001519	0.01648

After fitting ARMA-GARCH, we get the one-step forecasts on the log return and the volatility. And then the original covariance matrix $\mathbf{\Sigma}$ could be modified to refresh the diagonal elements, and $\mathbf{\Sigma}_{\boldsymbol{m}}$ is written as

$$\mathbf{\Sigma}_{\boldsymbol{m}} = \begin{pmatrix} 0.000154 & 0.000115 & 0.000171 & 0.000072 \\ 0.000115 & 0.0001467 & 0.000132 & 0.000250 \\ 0.000171 & 0.000132 & 0.000150 & 0.000006 \\ 0.000072 & 0.000250 & 0.000006 & 0.000272 \end{pmatrix}$$

And the return matrix $\boldsymbol{\mu}$ is given as

$$\boldsymbol{\mu} = \begin{pmatrix} 0.001530 \\ 0.001541 \\ 0.001677 \\ 0.001519 \end{pmatrix}$$

4.2 Apply time series forecast to the portfolio

Apply the modified $\mathbf{\Sigma}_{\boldsymbol{m}}$ and $\boldsymbol{\mu}$ into different portfolio strategies. We get the following outcomes.

Table 4.1. Weights distribution of each stock in 6 portfolios

$portfolio\backslash stock$	$google$	$apple$	$amazon$	$tencent$
RPP	0.2934	0.1778	0.2596	0.2691
$GMVP$	0.0203	0.9486	0	0.0310
$MSRP$	0.3502	0	0.2956	0.3542
$MVP(\lambda = 0.25)$	0	0	0.9538	0.0462
$MVP(\lambda = 0.5)$	0	0	0.7254	0.2746
$MVP(\lambda = 0.75)$	0	0	0.6493	0.3507

Table 4.2. Return and volatility 6 portfolios

$portfolio\backslash stock$	μ	σ
RPP	0.001567163	0.01317670
$GMVP$	0.001540094	0.01245172
$MSRP$	0.001569557	0.01248908
$MVP(\lambda = 0.25)$	0.001669694	0.01798040
$MVP(\lambda = 0.5)$	0.001633618	0.01466524
$MVP(\lambda = 0.75)$	0.001621593	0.01396523

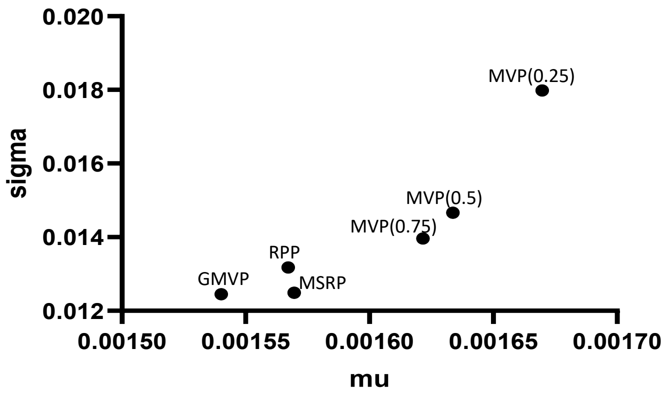


Figure 3.11: Distribution of the outcomes of 6 different portfolios

Figure 3.11 shows the distribution of return and volatility of 6 portfolios. It can be observed that, with the increasing of the μ, the σ would also increase, which means higher return brings higher risk.

The process of applying time series forecasts on portfolio construction can also be dynamic. Since this project uses daily log return and makes one-step-ahead forecast, the portfolio could be adjusted on each trading day based on the stock performance of the past several trading days.

The plots of the weight allocations would be in Appendix A for comparison and illustration.

Chapter 5
Limitations and Reflection

The whole logic and explanation of the project has been finished. Recall on the whole process, starting from data transformation, to the basic time series modelling (AR, MA, ARMA), to the advanced time series modelling (ARCH, GARCH, ARMA-GARCH), then to portfolio optimization. The final step is to apply the ARMA-GARCH modelling forecasts to the construction of the portfolio.

The fitting process are successful with the outcomes including the weight allocation of the stocks indifferent portfolio, the return and volatility of each portfolio, all of which could be extremely useful referencing tools for investors to make investment decisions.

Moreover, the idea and logic could be replicated and apply to most of the stocks, instead of just constrain to the certain stocks that has been modeled in this project. And it could also be dynamic within the changes of the stock market and refresh the portfolio allocation weights accordingly.

However, there are still room for improvements for the project.

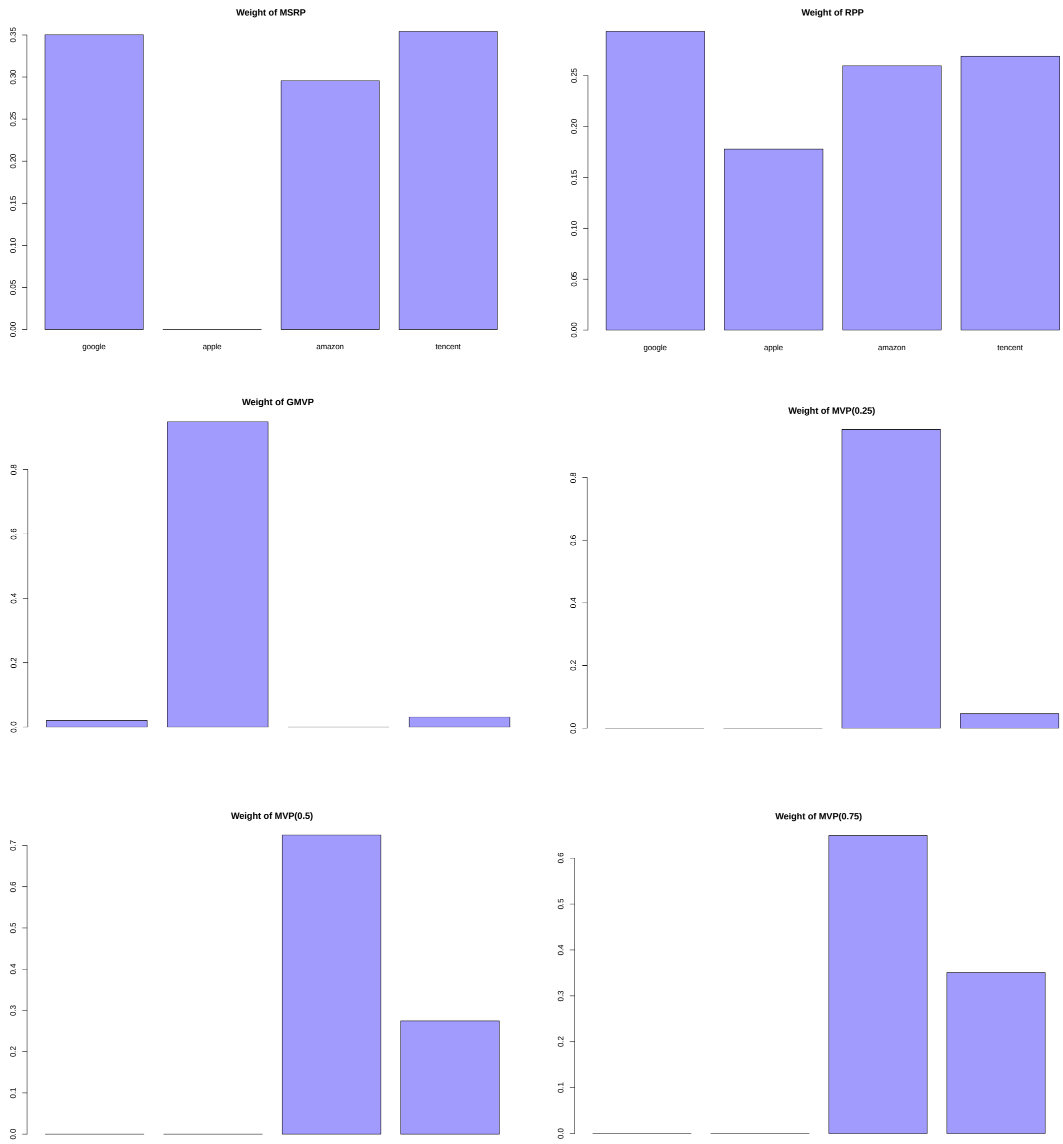
For example, in this project, we stop at ARMA-GARCH model, although the model is good enough to generalize the trend and volatility of time series, more advanced time series such as EGARCH, QGARCH, GJR-GARCH models could be studied and further applied under the design of the project. So, these advanced models could be my next-stage research topics.

Also, this project just simply adopted the classic portfolio strategies raised by the other scientists, without coming up with individual creative idea and insights of portfolio optimization. So, in the future research and study, a more dynamic and smarter portfolio with outstanding performances could be constructed. This may require advanced tools such as machine-leaning and artificial intelligence.

To state at the end of the project report, the financial market is too unpredictable and complicated to make completely accurate forecasts. One reason for its complexity is because its information and data are so massive that any current statistical or financial model with computer science tools have limited ability to intake and draw the whole picture. Another reason is that the market is constructed by people, while different individuals have different minds, different ideas, different preferences, different characters, all of these collide and compete with others which mutually effect the market. So apart from trying to use quantitative tools to predict the market, we shall also get down to the ground to observe the human nature. Only in this way can we get a deep and comprehensive understanding of the market operation disciplines.

Appendix A

Figure 3.12: Weight allocation based on the ARMA-GARCH fitting of 6 different portfolios



Appendix B

Algorithm of auto.arima()

Hyndman-Khandakar algorithm for automatic ARIMA modelling

1. The number of differences $0 \leq d \leq 2$ is determined using repeated KPSS tests.
2. The values of p and q are then chosen by minimising the AICc after differencing the data d times. Rather than considering every possible combination of p and q, the algorithm uses a stepwise search to traverse the model space.
   a. Four initial models are fitted:
      - ARIMA$(0, d, 0)$,
      - ARIMA$(2, d, 2)$,
      - ARIMA$(1, d, 0)$,
      - ARIMA$(0, d, 1)$.

      A constant is included unless
      $d = 2$. If
      $d \leq 1$, an additional model is also fitted:
      - ARIMA$(0, d, 0)$ without a constant.

   b. The best model (with the smallest AICc value) fitted in step (a) is set to be the "current model."
   c. Variations on the current model are considered:
      - vary p and/or q from the current model by ± 1;
      - include/exclude c from the current model.

      The best model considered so far (either the current model or one of these variations) becomes the new current model.
   d. Repeat Step 2(c) until no lower AICc can be found.

References

Vinícius, Z., & Palomar, D. P. (2019, October 5). Fast Design of Risk Parity Portfolios. https://cran.r-project.org/web/packages/riskParityPortfolio/vignettes/RiskParityPortfolio.html.

Back-testing and reporting of portfolio strategies. (2016). *Financial Risk Modelling and Portfolio Optimization with R*, 406–410. https://doi.org/10.1002/9781119119692.app3

Bayesian Approach of the Unit Root Test. (n.d.). *SpringerReference*. https://doi.org/10.1007/springerreference_205219

Best, M. J. (2010). *Portfolio optimization*. Chapman & Hall.

Bloss, M., Ernst, D., Häcker, J., & Sörensen, D. (2012). *Financial engineering*. Oldenbourg.

Box-Cox Transformation. (n.d.). *SpringerReference*. https://doi.org/10.1007/springerreference_205237

Dayal, V. (2015). *An Introduction to R for Quantitative Economics Graphing, Simulating and Computing*. Springer India.

Financial Analysis. (2015). *Analysis of Financial Statements*, 99–145. https://doi.org/10.1002/9781119203513.ch4

Financial Applications. (2010). *GARCH Models*, 311–340. https://doi.org/10.1002/9780470670057.ch12

The Great Subprime Meltdown of 2007. (2011). *Subprime Mortgage Credit Derivatives*, 293–318. https://doi.org/10.1002/9781118267165.ch13

Guo, X. (2017). *Quantitative Trading: algorithms, analytics, data, models, optimization*. CRC.

Hannan, E. J. (1983). *Time series analysis*. Chapman and Hall.

Hannan, E. J. (1983). *Time series analysis*. Chapman and Hall.

Harvey, A. (1995). *Time series*. Elgar.

Horton, N. J., & Kleinman, K. (2015). *Using R and RStudio for data management, statistical analysis, and graphics*. CRC Press, Taylor & Francis Group.

Härdle, W., Kleinow, T., & Stahl, G. (2002). *Applied quantitative finance: theory and computational tools*. Springer.

Jianguo, W. (2016). Estimating Stock Market Volatility with Asymmetric Arch, GARCH and Expanded GARCH Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2869039

Lai, T. L., & Xing, H. (2011). *Statistical models and methods for financial markets*. Springer.

Logarithm Transformation of the Output Response Data for Optimization. (2011). *Taguchi Methods*, 413–431. https://doi.org/10.1115/1.859698.ch12

Mooney, C. Z. (2003). *Monte Carlo simulation*. Sage.

Mushtaq, R. (2011). Augmented Dickey Fuller Test. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1911068

Pfaff, B. (2016). *Financial risk modelling and portfolio optimization with R*. Wiley.

Pok, W. C. (2009). *Financial market*. McGraw-Hill (Malaysia).

Portfolio Analysis in R - University of Washington. (n.d.). http://faculty.washington.edu/ezivot/econ424/portfolioTheoryRpowerPoint.pdf.

Rachev, S. T. (2011). *Financial models with Lévy processes and volatility clustering*. Wiley.

Richardson, C. H., & Miller, I. L. (2011). *Financial Mathematics*. Nabu Press.

Severini, T. A. (2017). *Introduction to Statistical Methods for Financial Models*. CRC Press.

Shimizu, K. (2010). Parametric ARMA(p, q)- GARCH(r, s) Models. *Bootstrapping Stationary ARMA-GARCH Models*, 65–83. https://doi.org/10.1007/978-3-8348-9778-7_4

Taniguchi, M., Amano, T., Ogata, H., & Taniai, H. (2014). *Statistical Inference for Financial Engineering*. Springer International Publishing.

End