Portfolio Construction and Time Series Forecasting

---Analysis Based on Dow Jones Industry Index

Siyuan Yao;
Tongyue Liu;
Xinyao Guo;
Yike Peng;
Yuting Zhang
Columbia University In The City Of New York

.

I Introduction

The goal of this paper is to give guidance to investors to construct the optimal stock portfolio. There are many theoretical papers on portfolio construction available, however, our paper aims to use stock market data and statistical analysis as well as modeling skills to apply both theoretical knowledge and analytical method in the real world. To specify, analysis of this paper is based on Dow Jones Index Average. In order to do data analysis in detail, we use the daily historical price of the Dow Jones 30 companies in 2016. To calculate difference financial measurements, we choose 1.22% (one year LIBOR in 2016) as the risk-free rate. The implementation was achieved by statistical language R and EXCEL.

This paper will use the Sharpe Ratio criteria to find a stock that performs the best, then construct the optimal combination of stocks by utilizing regression analysis as well as Lagrangian multiplier to allocate weight based on the constraint of minimum volatility. To measure and validate the performance of the portfolio, we will use Sharpe Ratio, Sortino Ratio, Value at Risk and time series modeling to satisfy investors' different risk appetites.

II Sharpe Ratio Criteria

Our first step is to find a stock that performs the best. In order to quantify financial performance, we use Sharpe ratio to measure the average return in excess of the risk-free rate per unit of volatility. Since Sharpe ratio is the risk adjusted return, a large Sharpe ratio means a high expected return, which indicate the asset/portfolio performs well.

In order to make comparison among the 30 stocks, we firstly calculate the expected daily return r_{daily} for each stock and convert them to the expected yearly return r_{yearly}:

$$r_{yearly} = (1 + r_{daily})^{252} - 1$$

Similarly, then we calculate the daily standard deviation for each stock σ_{daily} and convert them to yearly standard deviation σ_{yearly}:

$$\sigma_{yearly} = \sigma_{daily} \cdot \sqrt{252}$$

Next, we calculate the Sharpe ratio for each stock based on the risk-free rate r_f=1.22%, r_{yearly} is the expected yearly return for each company as calculated before(one year LIBOR in 2016):

$$\text{Sharpe Ratio} = \frac{r_{yearly} - r_f}{\sigma_{yearly}}$$

After computing Sharpe ratios, we sort the values and pick the stock with the highest Sharpe ratio (UNH) as our base asset. The following portfolio analysis is based on the stock we choose in this step.

---

①: r_{yearly} is the expected yearly return for each company as calculated before.

②: UnitedHealth Group, an American managed health care company based in Minnetonka, Minnesota. It is 6th in the United States on the Fortune 500. https://en.wikipedia.org/wiki/UnitedHealth_Group. Accessed April 24, 2017.

III Regression Analysis

1. Algorithm

Based on the chosen stock UNH before, we need to find other most suitable stocks from Dow Jones. So our main goal is to select the most irrelevant stocks of UNH and try to make them as an ideal portfolio based on the portfolio theory. Irrelevance means less risk, people tend to choose the safest one as well as making profits.

What we done to realize the goal including two steps, multiple regression and correlation test. Both of them are constructed for picking out the portfolio components.

Let Y be the response variable and $X_1, \dots, X_p$ be the predictor variables. Y_i and $X_{i,1}, \dots, X_{i,p}$ are the values of these variables for the ith observation. The regression modeling investigates how Y is related to $X_1, \dots, X_p$, estimates the conditional expectation of Y given $X_1, \dots, X_p$, and prediction of future Y values when the corresponding values of $X_1, \dots, X_p$ are already available. The multiple linear regression model rating Y to the predictor variables is $Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon =$ where ϵ_i is the random noise. β_0 is the intercept, it is the expected value of $YYii$ when all the Xi,j are zero. The coefficients of $\beta_1, \dots, \beta_p$ are the corresponding slopes of $X_{i,1}, \dots, X_{i,p}$. β_j Indicates every unit change in the expected value of Yi when $X_{i,j}$ changes on unit.

The multiple linear regression model is $Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$ We use the multiple linear regression model in R to find out the pair of assets with the smallest correlation. We firstly test the asset with the largest return with the other 29 assets and take into account the most correlated one, and test this new asset with the rest 28 assets and so forth. In the end, the last asset left is the least correlated to the first asset.

We do the stepwise regression which UNH is set as Y and the other 29 stocks as variables. The result can be showed in the following chart:

We do the stepwise regression which UNH is set as Y and the other 29 stocks as variables. The result can be showed in the appendix.

What showed above is an example of the regression process, we kept repeat the step for several times until all the variables left are not significant.

2. Correlation Test

We got 22 stocks left. In order to pick out our preferred stocks, we decide to do correlation test on these 22 stocks. Those who have lower correlated coefficients are the ones we need.

We have tried between 1 to 9 components, that means we chose the less relevant components based on the result, and compute the best portfolio measurements.

R codes can be found in the appendix.

Following are the example results of regression and correlation test:

Results of Correlation test

KO	0.236	JPM	0.458
PG	0.198	DO	0.281
MCD	0.318	VZ	0.149
MSFT	0.33	WMT	0.196
JNJ	0.228	NKE	0.284
CSCO	0.278	CVX	0.249
DIS	0.341	XOM	0.207
HD	0.374	AAPL	0.311
UTX	0.363	V	0.355
MMM	0.349	GS	0.459
INTC	0.39	CAT	0.286

IV Optimal combination calculation

1.Global minimum variance portfolio

One major problem, after we analyzed the multi-variance correlations matrix from candidate stocks, is how to set weight on a certain fixed number of assets. Recall our initial target is to figure out an optimal portfolio, a combination which could attain the global minimum risk. That is what we usually called as global minimum variance portfolio.

Therefore, this problem can be simplified as a process of calculating global minimum variance(risk) function. When we deal with weight allocation for multi-asset, especially for large-size portfolio, one commonly used method is to calculate the global minimum based on matrix algebra.

Take a simple example. Consider we invest a three-risky-asset portfolio with asset denoted as A, B, and C. Let the weight assigned on each stock be w_A , w_B , w_C . Then the global minimum risk of this portfolio can be expressed as:

$$\min_{w_A,w_B,w_c} \sigma_{p,w}^2 = w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + w_C^2\sigma_C^2 + 2w_Aw_B\sigma_{AB} + 2w_Aw_C\sigma_{AC} + 2w_Bw_C\sigma_{BC}$$

$$(w_A + w_B + w_c = 1)$$

2.Lagrangian multiplier

Lagrangian method is a commonly used approach when finding global minimum. The method can be implemented when there are extra constraints on parameters we tried to calculate. Since it is obvious that for any set of w_A , w_B , w_C , the sum of them always equals to one. With this equality constraints, we can here use Lagrangian multiplier to process the global minimum expression into a new Lagrangian function, in which we combine the original risk minimum function with the constraints function multiplied by a parameter λ:

$$L(w_A, w_B, w_c, \lambda) = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + w_C^2 \sigma_C^2 + 2 w_A w_B \sigma_{AB} + 2 w_B w_C \sigma_{BC} + \lambda (w_A + w_B + w_c - 1)$$

Then based on this Lagrangian function, we can calculate the global minimum with its first order gradient in term of each parameter:

$$0 = \frac{\partial L}{\partial w_A} = 2 w_A \sigma_A^2 + 2 w_B \sigma_{AB} + 2 w_C \sigma_{AC} + \lambda$$

$$0 = \frac{\partial L}{\partial w_B} = 2 w_B \sigma_B^2 + 2 w_A \sigma_{AB} + 2 w_C \sigma_{BC} + \lambda$$

$$0 = \frac{\partial L}{\partial w_C} = 2 w_C \sigma_C^2 + 2 w_A \sigma_{AC} + 2 w_B \sigma_{BC} + \lambda$$

$$0 = \frac{\partial L}{\partial \lambda} = w_A + w_B + w_c - 1$$

3.Matrix algebra

To simplify the process of calculation we transform gradient equality into matrix and get final result:

$$\begin{pmatrix} 2\sigma_A^2 & 2\sigma_{AB} & 2\sigma_{AC} & 1 \\ 2\sigma_{AB} & 2\sigma_B^2 & 2\sigma_{BC} & 1 \\ 2\sigma_{AC} & 2\sigma_{BC} & 2\sigma_C^2 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} w_A \\ w_B \\ w_c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

$$c(w_A, w_B, w_c) = c(0.4521, 0.2271, 0.3208)$$

By applying steps similar to those given above, we calculated the optimal weight for portfolios in size of 1 to 10 stocks separately. Here we show the optimal weights for 3-asset portfolio, 5-asset portfolio and 10-asset portfolio, and then plot the efficient frontier for tangency portfolio of each size(here we only show the plot for 10-asset case). Comparing the performance of optimal invests from different size level we are able to pick out the optimal size.

Three-asset portfolio weight

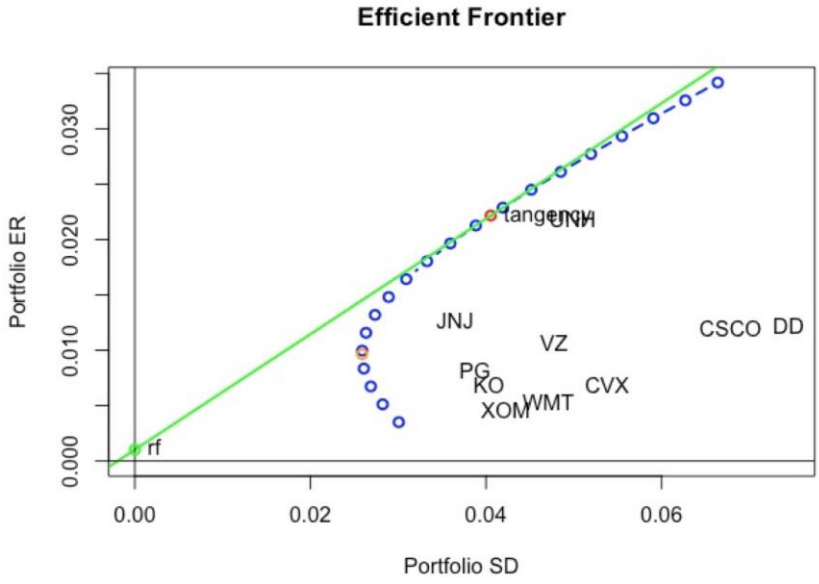
Stock	VZ	WMT	UNH
Weight	45.21%	22.71%	32.08%

Five-asset portfolio weight

Stock	VZ	WMT	UNH	PG	XOM
Weight	21.62%	12.33%	21.20%	33.78%	11.07%

Ten-asset portfolio weight

Stock	VZ	WMT	UNH	PG	XOM	KO	JNJ	CSCO	DD	CVX
Weight	10.6%	8.8%	13.41%	11.73%	9.53%	20.4%	32.49%	2.95%	5.93%	9.94%



IV Performance Measurement

In the first step choosing stocks, we try our best to choose stocks that has lowest correlation which means the dependence was very weak, thus we default the stocks chosen in portfolio were almost independent.

1. Sharpe Ratio

Sharpe Ratio calculates the average return earned in excess of the risk-free rate per unit of volatility. The greater the value of the Sharpe Ratio, the better the performance of the risk-adjusted return. The results are shown in Table below.

	3 stocks portfolio	5 stocks portfolio	10 stocks portfolio
Return	0.2785	0.2235	0.1607
STD	0.1234	0.1102	0.0986
Sharpe Ratio	2.16	1.92	1.51

From the form, we saw that the Sharpe ratio of the 3 stocks portfolio was the biggest with the value 2.16. Therefore, we made a preliminary conclusion that the 3 stocks portfolio with mean 0.2785 and standard deviation 0.1234 was the optimal among all the others.

2. Sortino Ratio

The Sortino ratio is a variation of the Sharpe ratio that the excess return over the risk-free rate divided by the downside deviation(the asset's standard deviation of negative asset returns), it can help investors to assess risk in the return to "bad" volatility(Volatility caused by negative returns is considered bad or undesirable by an investor).

From our dataset, we found there were nearly half of return are negative, in this case for investors who are more sensitive to asset value downside, sortino ratio is an important measurement tool. The results are shown in Table below.

	3 stocks portfolio	5 stocks portfolio	10 stocks portfolio
Return	0.2785	0.2235	0.1607
downside deviation	0.081	0.0685	0.06497
Sortino Ratio	3.300	3.08	2.29

Three stocks portfolio has the highest sortino ratio so it manifested the same conclusion that the 3 stocks portfolio was the best among the all portfolios and made it more evident.

3. Value at Risk and Conditional Value at Risk

In financial risk management, Value at risk (VaR) and Conditional Value at Risk (CVaR) are two risk assessment technique used to reduce the probability that a portfolio will incur large losses.VaR is used to measure the level of financial risk within a investment portfolio over a specific time frame,CVaR is used to assess the likelihood that a specific loss will exceed the VaR at a specific confidence level. We used them to see after investing $ 1 million, the loss value based on the confidence level of 95% and 97.5%. The output is shown in Table below.

Annual	3 stocks portfolio	5 stocks portfolio	10 stocks portfolio
VaR at 95%	-11153.43	-10643.3	-9829.86
VaR at 97.5%	-15062.80	-14441.99	-11513.45
CVaR at 95%	-16113.29	-14426.00	-12803.63
CVaR at 97.5%	-18851.25	-15944.02	-14523.97

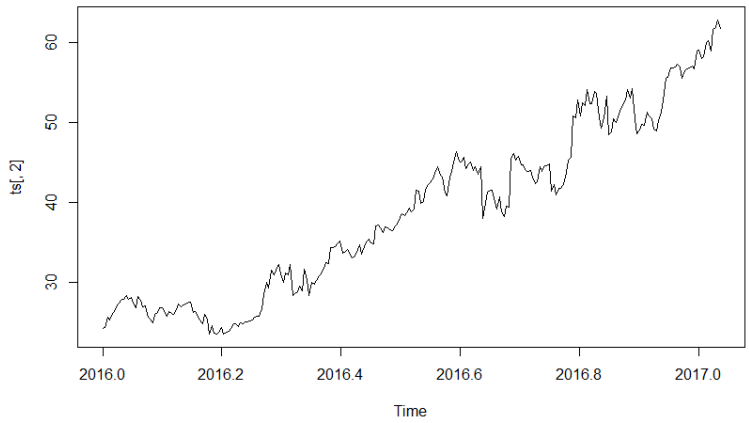
From the VaR, it shows that there is a 5% chance that the 3 stocks portfolio will fall in value by more than $ 11153.43 and 2.5% chance that the 3 stocks portfolio will fall in value by more than $ 15062.80 based on a $ 1 million investment, which is largest loss among these different portfolios. CVaR shows that 3-stock portfolio has the highest average of the losses that occur beyond the Var cutoff point. In this case, we infer a risk-averse investor will choose 10 stocks portfolio as their optimal portfolio.

In a word, according to various investor’s demand we should choose different portfolio as optimal.

V Time Series Models

1.ARIMA Model Analysis

a. Plot Portfolio Price in time series



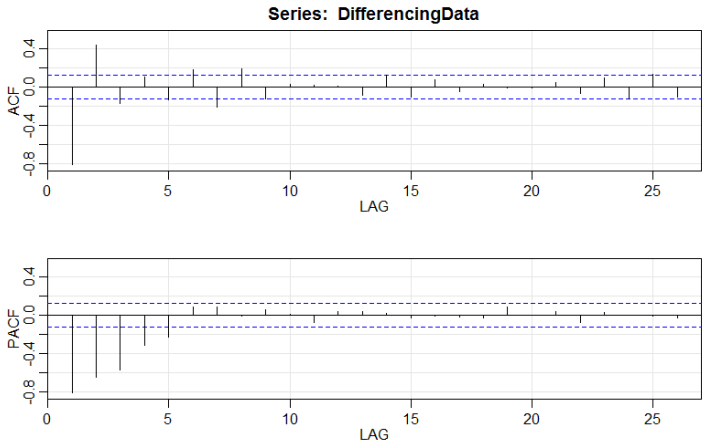
b. Differencing a Transformed Time Series

ARIMA models are defined for stationary time series. Therefore, if we start off with a non-stationary time series, we will first need to 'difference' the time series until we obtain a stationary time series. Log smoothing methods are useful for making forecasts and have no need to make assumptions about the correlations between successive values of the time series.

While log smoothing methods do not make any assumptions about correlations between successive values of the time series, in some cases we can make a better predictive model by taking correlations in the data into account. Autoregressive Integrated Moving Average (ARIMA) models include an explicit statistical model for the irregular component of a time series, which allows for non-zero autocorrelations in the irregular component.

2. Selecting a Candidate ARIMA Model

Since we have transformed it to a stationary time series by differencing d times, the next step is to select the appropriate ARIMA model. In other words, we have to find the values of most appropriate values of p and q for an ARIMA(p,d,q) model. To do this, we usually need to examine the correlogram(ACF) and partial correlogram(PACF) of the stationary time series.

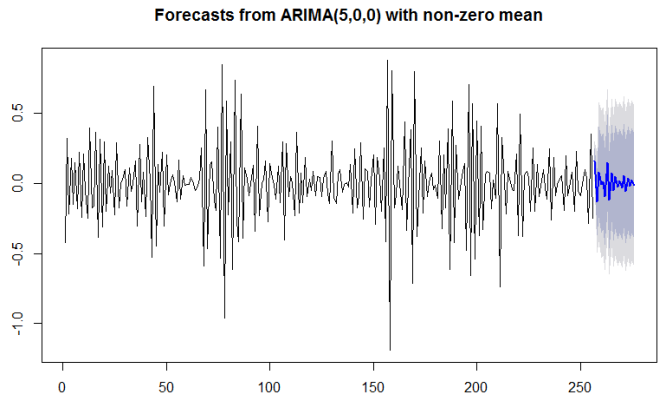


From the correlogram, we can observe that the autocorrelations for lags 1, 2 and 3 exceed the significance bounds, and the autocorrelations tail off to zero after lag 3. The autocorrelations for lags 1, 3 are negative in magnitude(lag 1: -0.8, lag 2: 0.412, lag 3: -0.019). However, the autocorrelation for lags 6, 7 and 8 slightly exceeds the significance bounds as well, it is likely that this is due to chance. The autocorrelations for rest lags do not exceed the significance bounds, and we would expect 1 in 20 lags to exceed the 95% significance bounds by chance alone.

From the partial auto correlogram, we see that the correlogram tails off to zero after lag 5, and the partial correlogram is zero after lag 5, the following ARMA models are possible for the time series:

·An ARMA(0,3) model, since the autocorrelogram is not clearly zero after lag 3, the model maybe not suitable for our data.

·An ARIMA(5,1,1) mixed model, since the correlogram and partial correlogram tail off to zero, but the partial correlogram perhaps tails off too abruptly for this model to be appropriate.



The forecasts are shown as a blue line, with the 80% prediction intervals as an grey shaded area, and the 95% prediction intervals as a dark grey shaded area.

Conclusion

In conclusion, to construct an optimal stock portfolio based on the Dow Jones 30 companies, our first step is using the Sharpe Ratio measurement to pick UnitedHealth Group (UNH) and include this stock in our portfolio.

Regression and Correlation test are used to select the best suitable components in the portfolio, selections are based on the result of correlation test. We picked up different numbers of portfolios and calculate their weights separately to get our best portfolios.

The portfolio has a relatively high return compared to the historical data and price has a seasonal trend. In general case, price always go down in December.

There is a tradeoff between high return and low volatility. For risk-averse investor, it is better to choose 10-stock portfolio based on VaR analysis. Otherwise, choose 3-stock portfolio based on Sharpe Ratio and Sortino ratio with help of ARIMA model to predict.

Furthermore, our research is based on 30 representative stocks selected in Dow Jones Index Average, if we choose different dataset the optimal portfolio may be change, we will do future research later. While we believe our portfolio construction method can be generalized and applied in the greater scope.

Appendix

1. Regression Model Summary

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0010448  0.0006563   1.592  0.11279
KO           0.0979748  0.1060471   0.924  0.35653
PG          -0.0839880  0.1050437  -0.800  0.42481
PFE          0.1864730  0.0689549   2.704  0.00736 **
MCD          0.1350060  0.0882004   1.531  0.12724
MSFT        -0.0135691  0.0713949  -0.190  0.84943
JNJ         -0.1000078  0.1049971  -0.952  0.34187
CSCO        -0.0437180  0.0695351  -0.629  0.53017
DIS          0.0509395  0.0875086   0.582  0.56107
HD           0.0654828  0.0793765   0.825  0.41026
UTX          0.0718050  0.0867486   0.828  0.40869
MMM          0.0647955  0.1129414   0.574  0.56673
INTC         0.0972787  0.0706252   1.377  0.16975
JPM          0.0836245  0.0937894   0.892  0.37354
DD          -0.0067192  0.0596715  -0.113  0.91045
VZ          -0.0675788  0.0819973  -0.824  0.41071
WMT          0.0705543  0.0635000   1.111  0.26770
NKE          0.0152286  0.0579503   0.263  0.79295
CVX          0.0324006  0.0738022   0.439  0.66106
XOM         -0.0563753  0.0883822  -0.638  0.52421
AAPL         0.0867305  0.0541244   1.602  0.11045
V           -0.0178811  0.0737468  -0.242  0.80864
GS           0.1319616  0.0861685   1.531  0.12705
CAT         -0.0640613  0.0625117  -1.025  0.30655
Residual standard error: 0.0102 on 227 degrees of freedom
Multiple R-squared:  0.3392,          Adjusted R-squared:  0.2722
F-statistic: 5.065 on 23 and 227 DF,  p-value: 2.882e-11
```

2. Code

```
##########################Data Extraction
library(stockPortfolio)
ticker  <- c("KO", "PG",
"PFE","MCD","MRK","MSFT","JNJ","BA","CSCO","DIS","IBM","HD","UTX","MMM","INTC","JPM
","DD","UNH","VZ","WMT","NKE","CVX","XOM","GE","AAPL","V","TRV","GS","CAT","AXP")
returns=getReturns(ticker, "day", start="2016-01-01",end = "2016-12-31")
data<-returns$R
riskfree_rate=0.0122

##########################Sharpe and Sortino Ratio
sharpe_ratio<-rep(NA, length(ticker))
```

```
for (i in 1:length(ticker)) {
  sharpe_ratio[i]<-(((mean(returns$R[,i])+1)^252-1)-riskfree_rate)/(sd(returns$R[,i])*sqrt(252))}
names(sharpe_ratio)<-ticker
which.max(sharpe_ratio)
data1<-data.frame(data)

sortino_ratio<-rep(NA, ncol(data1))
negative_r<-matrix(NA,3548,ncol(data1))
sd<-rep(NA, ncol(data1))
riskfree_rate<-0.0122
for (i in 1:ncol(data1)) {
  negative_r[,i]<-data1[,i][data1<0]
  sd[i]<-sd(negative_r[,i][!is.na(data1[,i][data1<0])])
  sortino_ratio[i]<-(((mean(data1[,i])+1)^252-1)-riskfree_rate)/(sd[i]*sqrt(252))}
names(sortino_ratio)<-ticker
which.max(sortino_ratio)

############################Data Selection
library(broom)
UNH<-data1$UNH
data1<-data.frame(data)
fit <- lm(formula = data1$UNH ~., data=data1)
temp<-tidy(fit)$p.value
temp<-temp[-1]
data2<-data1[,-18]
temp.data<-data2[,temp>0.1]
temp.data<-data.frame(temp.data,UNH)

###################Linear Regression
fit1 <- lm(formula = temp.data$UNH ~., data=temp.data)
temp<-tidy(fit1)$p.value
temp<-temp[-1]
data2<-temp.data[,-24]
temp.data<-data2[,temp>0.1]
temp.data<-data.frame(temp.data,UNH)

############################Correlation Selection Based on Regression
n=3
d=temp.data
cor(d)
test<-cor(d)[23,]
num<-order(test)
d1<-d[,num[1:n-1]]
d1<-data.frame(d1,UNH)
```

```
############################Select Weight
sigma.mat=cov(d1)
top.mat = cbind(2*sigma.mat, rep(1, n))
bot.vec = c(rep(1, n), 0)
Am.mat = rbind(top.mat, bot.vec)
b.vec = c(rep(0, n), 1)
z.m.mat = solve(Am.mat)%*%b.vec
m.vec = z.m.mat[1:n,1]
m.vec

############################Time Series Preparation
library(forecast)
library(TTR)
library(astsa)
library(tseries)
full=getReturns(ticker, "day", start="2016-01-01",end = "2016-12-31")$full
price<-matrix(rep(0,30*252),ncol = 30)
for(i in 1:30){price[,i]<-full[[i]]$Adj.Close}
colnames(price)<-colnames(data1)
price0<-price[,c("UNH","VZ","WMT")]
price0<-data.frame(full[[1]]$Date,price0)
colnames(price0)[1] <-"Date"
price0<-price0[,-1]

#equalprice<-rowSums(price0)/10
#equalprice0<-data.frame(full[[1]]$Date,equalprice)
#colnames(equalprice0)<-c("Date","equalweightprice")

#data2<-equalprice0[order(equalprice0$Date,decreasing=F),]
#ts<-ts(data2, frequency=252, start=c(2016,1,4))
#par( mfrow = c(1,1) )
#plot.ts(ts[,2])

a=data.matrix(price0, rownames.force = NA)
b=data.matrix(m.vec, rownames.force = NA)
price0=a%*%b
wprice0<-data.frame(full[[1]]$Date,price0)
colnames(wprice0)<-c("Date","wprice0")
data2<-wprice0[order(wprice0$Date,decreasing=F),]
ts<-ts(data2)
par( mfrow = c(1,1) )
plot.ts(ts[,2])
```

```
###############ARIMA
data2<-diff(log(ts[,2]))
acf2(data2)
fit.arima<-arima(data2, order=c(9,1,9))
acf2(resid(fit.arima))
hist(resid(fit.arima))
summary(fit.arima)
fit.ar<-arima(data2, order=c(9,0,0))
fit.ma<-arima(data2, order=c(0,0,9))

############Prediction
forecast(fit.arima,h=5)
plot(forecast(fit.arima,h=5))
plot(forecast(fit.ar,h=5))
plot(forecast(fit.ma,h=5))

#############Model Comparison
fit.arima$aic
fit.ar$aic
fit.ma$aic
```

Bibliography

Li, Z. (2015). Statistical method for risk management and portfolio theory. Investopedia Staff, retrieved from http://www.investopedia.com/terms/v/var.asp