Equity Statistical Arbitrage

Xinyi Li

Introduction

Under the similar economic background, equities share the risk exposures, so they share the similar price movement. The equity price movements can be decomposed to major risk exposure, idiosyncratic risk and a slow price drift. We can assume idiosyncratic risk is mean-reverting and estimate the risk exposure of bundles of equities by principal components. If the difference between the theoretical price and real equity price exists, there will be some opportunities to arbitrage.

I. Specification

A. Universe

Our strategy analyzes the performance of equity statistical arbitrage in U.S. stock market. We collected totally 482 stocks which are existed between 1/1/2006 and 12/31/2013 from 8 different industries including utilities, healthcare, services, industrial goods, basic materials, consumer goods, technology and finance. We want the data to be sufficiently and comprehensive to support our strategy implementation. However, the survivorship biased exists in the data selection.

B. Date Range

Our data range is between 1/1/2006 and 12/31/2013. We think 8 years is a suitable period to implement our strategy. The data range includes the 2008 financial crisis which means we can analyze the strategy performance during the financial crisis. We can perform some risk management analysis such as VaR. Also, it is easy for us to check the correlation between the market and strategy during the market crash. We used 400 stocks for the principle components analysis. 25 target stocks are used for in-sample, while 15 stocks are used for out-of - sample.

C. Data Sources

Our data is downloaded from Yahoo finance, since Yahoo finance is an authoritative, accurate and accessible financial source website.

D. Signal Generation

Computing the Signal

1) Analyze market with PCA to design risk bundles

$$\sum_{j=1}^{N} \beta_{ij} R_j$$

2) Regress the target returns against the risk bundles

$$R_n^S = \beta_0 + \beta R_n + \varepsilon_n \ , \mathrm{n} = 1,2,\dots,60$$

3) Analyze the residuals as an AR(1) process. Estimate the a, b, Var (ζ) below.

$$X_{n+1} = \varphi X_n + \zeta_{n+1}, n = 1,2,\dots,59$$

4) Extract the walk parameters and observe the signal. Note that m bar below is the average position of m across many stocks.

$$\hat{\sigma} = \sqrt{\mathrm{Variance}(\zeta)}\mathrm{m} = \beta_0 + \beta R_n, \mathrm{s} = \frac{X_{60} - m}{\hat{\sigma}}$$

5) Trading the signal: Use the mean reversion of the idiosyncratic risk to sell short when s is high and go long when s is low.

E. Portfolio Construction

The portfolio is equally weighted for all target stocks. (25 target stocks for backtest; only the mean performance of the 25 are cared about) For each stock, we repeated the process given below:

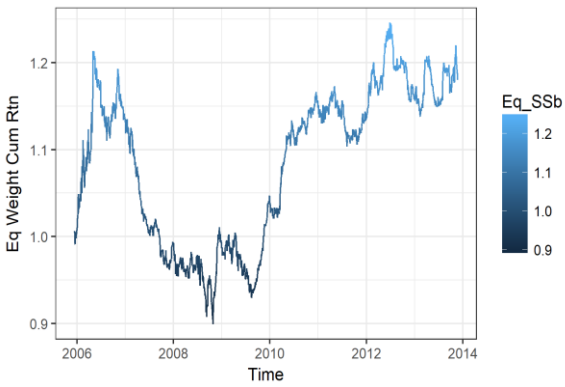
1) Representing the stocks return data on given dates and going back M+1 days as a matrix
2) Identify principle components; Adopt principle components covering 55% variance
3) Generate signal, according to steps in Part D
4) Go into long/short position according to the sign of signal: Long target stock short principle components or Long principle components short target stock with adjusted coefficient.

F. Execution

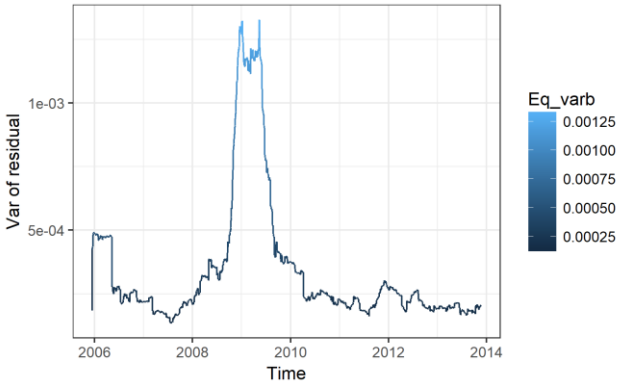
-Buy to open if $s_i < -1$
-Sell to open if $s_i > +1$
-Close short position if $s_i > +1.5, or\ s_i < -0.5$
-Close long position if $s_i > 0.75, or\ \ s_i < -2$
-Force to close position after holding 90 days if not meeting the above criterion

II. Implementation

A. PnL Graph



Plot-1 Back-test Cumulative Return



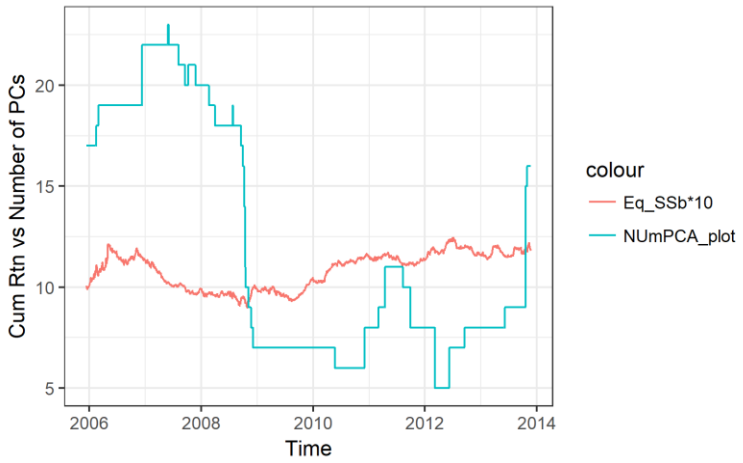
Plot-2 Back-test Variance of Residual

1) Sensibility

The strategy is based on the assumption that a set of statistical relationships, co-intergrations, will revert to their historical means. A drift from a historical relationship implies the market did not price a specific asset with a specific factor change, hence an arbitrage opportunity exists.

2) Efficacy

The strategy was lost in around 2008. 2009 the rate of return has risen rapidly and maintained at a certain level. From Plot-1, the strategy get the higher cumulative return during 2006-2007 and 2012-2013. The rate of cumulative return continues to oscillate over the past 2008 to 2010 years. That is to say, ignoring the impact of the financial crisis, the cumulative yield of the strategy is good. This is also reflected in Plot-3. From Plot-2 we can see that the estimated variance of residuals increased a lot after 2008 which could be suggested by financial crisis. With the big variance of residual and small number of principal components, the performance should be better than other times suggested by common papers. While in our implementation, in the period of 2008-2009, we do see a small return from previous downtrend, but that's far from our expectation.



Plot-3 Cumulative Return vs Number of PCs

3) Adjustments
   a) Trigger values for open / close positions are adjusted according to signals.
   b) Minimum 5 Principal Components are guaranteed.
4) Costs
   We ignored the Transaction costs in back-test. It will be considered in refinement.
   Also, financing costs and other execution costs are ignored in our strategy.

B. Stats

Table-1 Stats

Annualized Return	2.09%	Skewness	0.327
Annualized Risk	5.88%	Kurtosis	5.259
Max Ann. Return	23.06%	1-day 95% VaR	0.57%
Min Ann. Return	-11.89%	1-month 95% VaR	2.85%
Sharpe Ratio	0.356	Average |Signal|	14.26

The annualized return is 2.09% and Sharpe ratio is only 0.356, which is much lower than our expectation. Also, the signal is larger than our expectation. We need to either perform more precise regression or bring more stocks in to PCA.

C. Difficulties

1) As indicated from the table, the average |Signal| is 14.26, which is too high for the trigger values for open/close position. We need to find suitable trigger values; otherwise the holding period will be extremely short.
2) With large absolute value of signal, the estimation of the variance of regression residuals might be improved.
3) Transaction cost needs to be considered.

III. Refinements

A. Implemented

1) Transaction Costs
   In the first implementation backtest, we didn't consider the transaction costs. Actually, transaction costs occur every time we open a new position (i.e. when we short our target stocks or short those principal components stocks). To make our model more realistic, we add a fixed proportion of transaction costs to our strategy.

2) Time Series Modeling

The paper chose AR (1) model to get the coefficient and variance of the model to calculate the signal. However, sometimes the AR (1) may not have a good prediction power, that is, there may be some other time series model that fits our return data better. To avoid the bias that may be caused by a poor-fitted model, we directly extract the sample autocovariance at lag 0 (h = 0) to calculate the signal.

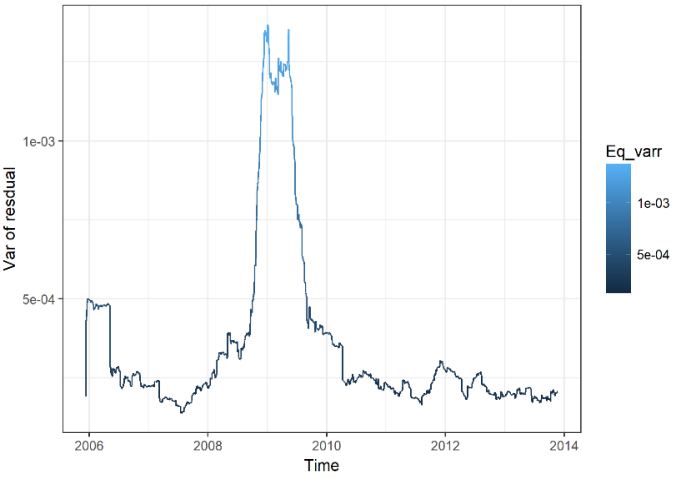
3) Weighting Method

For the first implementation and backtest, we used equally weighting to construct our portfolio. For those traditional portfolio construction processes, we usually consider putting higher weight on those stocks that have some better metrics than others, such as higher momentum and better earnings record. However, here the way we make money is way different to those traditional ones: we are trying to seize the inequality between the theoretical return and the actual return of our target stocks through the regression on PCA and make profit by either longing or shorting our target stocks, which means there's actually not any significant relationship between the recent performance of stocks and the return of our strategy. Hence, we decide not to keep using equally weighting.

4) PCA/Regression Frequency

For this strategy, PCA and regression are combined to build model and calculate the trading signal. We redo the PCA every 180 days if there's no position enrolled. However, we don't know exactly how long one PCA model can hold to be effective and can keep explaining enough variance of our model. For the future research, improvements related to this part may be added to the strategy to ensure the effectiveness of our PCA model.

Plot-4 Refinement Cumulative Return



Plot-5 Refinement Variance of Residual

Table-2 Refinement Stats

Annualized Return	1.70%	Skewness	0.302
Annualized Risk	5.88%	Kurtosis	5.242
Max Ann. Return	22.43%	1-day 95% VaR	0.57%
Min Ann. Return	-12.30%	1-month 95% VaR	2.40%
Sharpe Ratio	0.290	Average |Signal|	16.57

After refinement, with the consideration of transaction cost, the annualized return and Sharpe ratio is even lower because of transaction cost, which will generally cut off 1% return annually. The annualized risk stays the same and the signal is still large.

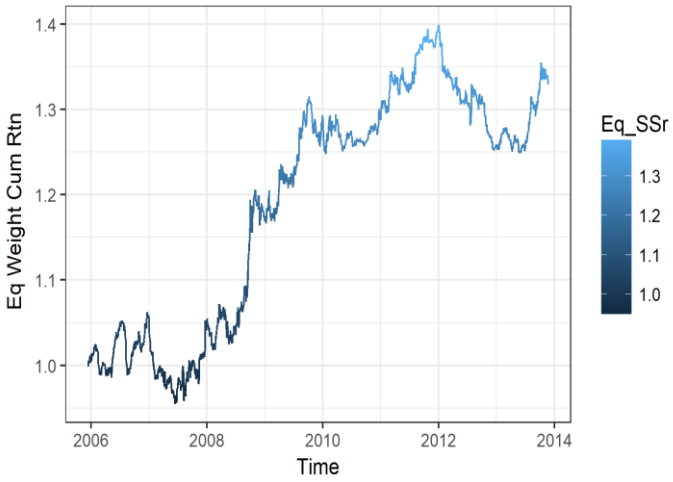
Generally speaking, using sample ACVF at lag 0 does not change a lot from the back-test, where AR(1) model is used. Based on these analysis, we don't believe the method to estimate variance of residuals from regression can influence general strategy performance. We will also plot similar plot in the Out-of-sample test to verify this claim.

B. Proposals

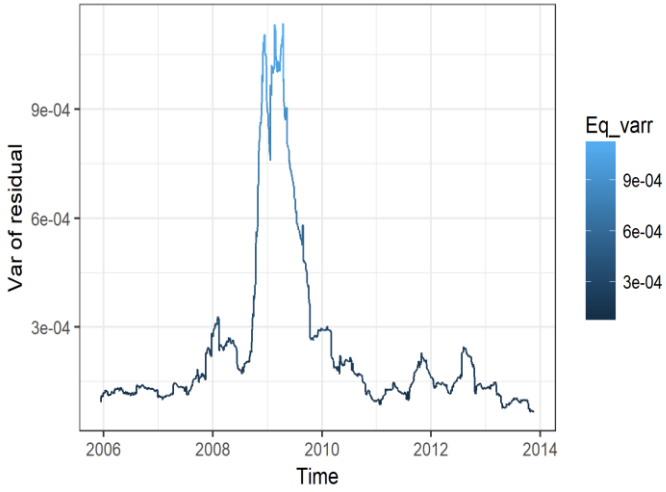
For our implementation, 400 stocks are gathered to conduct the Principal Components Analysis. For future research, researchers can probably use more stocks to do the PCA (e.g. 800 or 1000 stocks) for a more comprehensive explanation of our risk-bundles.
In addition, as what's shown in our signal calculation, some signals are extremely large, which may be considered as outliers. Finding a threshold to eliminate those potential outliers may be a reasonable choice for getting a more reliable model.
Finally, 60 days are chosen to do the regression on principal components (as shown in the paper). Probably a longer regression period could provide a better estimation on the way those risk-bundles make contribution to the stock return.

IV. Conclusion

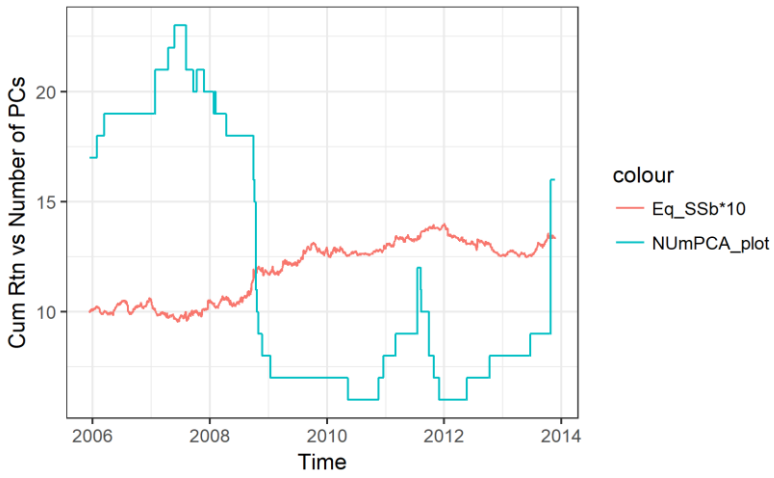
A. Out of Sample Test(s)



Plot-6 Out of Sample Cumulative Return



Plot-7 Out of Sample Variance of Residual



Plot-8 Cumulative Return vs Number of PCs

Table-3 Out of Sample Stats

Annualized Return	3.62%	Skewness	0.945
Annualized Risk	5.91%	Kurtosis	13.15
Max Ann. Return	13.91%	1-day 95% VaR	0.54%
Min Ann. Return	-3.44%	1-month 95% VaR	1.87%
Sharpe Ratio	0.614	Average |Signal|	8.59

As we can see from the graphs, with a different test samples, the difference of return and risk could be significant to some extent, which can be seen from these metrics like annualized return and Value at Risk. Meanwhile, some patterns still hold for out-of-sample test. Firstly, strategy performs better when less Principal Components are used and secondly, the signals of our portfolio are still to large even though they reduced by half for this test sample.

B. Trading Recommendation

Overall, equity statistical arbitrage, as a highly mathematical and theoretical trading strategy, did provide a high return during some period but, the performance is not stable enough to convince investors rely too much on it. Moreover, modeling risk may lead to a calculation bias since there're lots of inputs variables needing a timely renew (PCA/Regression) to guarantee the validity of the algorithm. In a word, if investors have strong confidence on the explanation power of PCA and regression, and also on the effectiveness of the algorithm, this strategy could be a quite valuable alternative.

V. Appendix

Code for Equity Statistical Arbitrage

```
# dataclean function
  dataclean <- function(filepath){
  library(plyr)
  # read .csv in filepath
  readwks <- function(file) {
    filenames <- list.files(path = file, pattern = "*.csv", full.names = TRUE)
    ldply(filenames, .fun = read.csv)
  }
  a <- readwks(filepath)
  #count stock number
  num_stocks <- length(list.files(path = filepath, pattern = "*.csv"))
  # output stock name
  name_stocks.csv <- list.files(path = filepath, pattern = "*.csv")
  name_stocks <- substr(basename(name_stocks.csv), 1, nchar(basename(name_stocks.csv)) - 4)
  # construct data output matrix with date and stock name
  b <- matrix(a[,"Adj.Close"], nrow = nrow(a)/num_stocks, ncol = num_stocks)
  date <- a[1:(nrow(a)/num_stocks),'Date']
  rownames(b) <- date
  colnames(b) <- name_stocks
  return(b)
}
library(quantmod)
data=dataclean("~/test/test")

for (i in 1:length(data[1,])){
```

```
  data[,i] <- dailyReturn(data[,i])  # Daily return
}

PCA <- function(Variance,data) {
  pcax <- prcomp(data)  # PCA
  vars <- apply(pcax$x,2,var)
  props <- vars / sum(vars)  # Var propotion of each PC
  cumprops <- cumsum(props)  # Cum Var propotion
  k <- 5
  for (i in 5:(length(cumprops)-1))
    if ((cumprops[i] < Variance) & ((cumprops[i+1] > Variance)))
      k <- i+1
  w=pcax$rotation[,1:k]
  return(list(PCw = w , K = max(k,5)))  # pick PCs makes %Variance variance
}

Regression <- function(i,t,x,PCw,data) { # regression target port. aginst Picked PCs previously
  Rtn <- data[(i-t):(i-1),]%*%PCw
  Reg <- lm(x[(i-t):(i-1)] ~ Rtn)
  return(list(res = Reg$residuals , coef = Reg$coefficients))  # Return residuals for Time Series Analysis
}

ACVF <- function(res) {    # get ACVF at h = 0
  library(itsmr)
  gamma_0 <- acvf(res, h = 0)
  return(gamma_0)
}

Signal_cal <- function(x_i,gamma_0,i,beta,beta0,Rtn_i) {  #Signal Calculation
  m <- beta0+Rtn_i%*%beta
  signal <- (x_i-m)/sqrt(gamma_0)
  return(signal)
}

Dailyrtn <- function(i,PCw,beta,beta0,y,data,sgn) {   # CAlculate everyday return
  m <- (data[i,]%*%PCw)%*%beta+beta0    # m: model return, x: target port. return
  Dr <- sgn*(y[i]-m)       # Daily return
  return(Dr)
}

startdate <- 2000   # start date   # same period for back test and refinement!
enddate <- 4000    # end date      # Time: 2006/1/1 - 2013/12/31
period_PCA <- 500  # Time frame for PCA (using 500 days' data before day i)
period_reg <- 60   # Time frame for regression (using 60 records before day i)
var_prop <- 0.55   # PCs with 60% variance
trans <- 0.0001   # transcation cost

FLAG1 <- FALSE    # identify if we have position
LONG <- FALSE     # identify Long/short of the position
FLAG2 <- TRUE     # identify if we need to redo PCA
```

```
data_I <- data_PCA  # Stocks we use to build model

NumPCA <- rep(0,5000)   # Record the number of PCs

SS <- rep(1,5000)
Dr <- rep(0,5000)
backtest <- data_b    # target stock for back test
refinement <- data_r   # target stock for refinement

SSb <- matrix(1,ncol = length(data_b[1,]),nrow = 5000) # record cumulative return
Drb <- matrix(0,ncol = length(data_b[1,]),nrow = 5000) # record daily return
var1b <- matrix(0,ncol = length(data_b[1,]),nrow = 5000)  # record estimate var of error for regression

SSr <- matrix(1,ncol = length(data_r[1,]),nrow = 5000) # record cumulative return
Drr <- matrix(0,ncol = length(data_r[1,]),nrow = 5000) # record daily return
var1r <- matrix(0,ncol = length(data_r[1,]),nrow = 5000)  # record estimate var of error for regression

for (j in 1:length(data_b[1,])){
  x <- data_b [,j]
  # x <- data_r[,j]
  for (i in startdate:enddate){
    if (FLAG2 == TRUE){
      PCw <- PCA(var_prop,data_I[(i-period_PCA):(i-1),])[[1]]  # initalize PCA every 180 days/after closing out
position
      k <- PCA(var_prop,data_I[(i-period_PCA):(i-1),])[[2]]
      t0 <- i
    }
    NumPCA[i] <- k
    FLAG2 <- FALSE
    if ((i-t0) >= 180)        # if no position enrolled in previous 180 days, redo PCA next time
      FLAG2 <- TRUE

    if (FLAG1 == FALSE){      # we donlt have position now

      res <- Regression(i,period_reg,x,PCw,data_I)[[1]]   # indentify residule of regression
      coef <- Regression(i,period_reg,x,PCw,data_I)[[2]]
      beta0 <- coef[1]    # intercept of regression
      beta <- coef[2:length(coef)]   # beta of each PC

      gamma_0 <- ACVF(res) # get acvf at h = 0

      FLAG <- TRUE            # Regression with large VAR of Error
      if (sqrt(gamma_0) > 0.33){
        FLAG <- FALSE
      }

      Rtn_i <- data_I[i,]%*%PCw    # PCs' return
      signal <- Signal_cal(x[i],gamma_0,i,beta,beta0,Rtn_i)
      if ((abs(signal) > 2) & (FLAG == TRUE)){
        FLAG1 <- TRUE
        t1 <- i
```

```
    if (signal>2){   # Judge Long or Short
      LONG <- FALSE
      sgn <- -1      # For daily return calculation//signal weighting
    }
    else{
      sgn <- 1     # For daily return calculation//signal weighting
      LONG <- TRUE
    }
  }

  Drb[i,j] <- 0   # for backtest, no transaction cost
  SSb[i,j] <- SSb[i-1,j]    # record simulated return

  if (FLAG1 == TRUE){  # for refinement, with transaction cost
    if (LONG == TRUE){  # open long position, sell PCs will cause transaction cost
      Drb[i,j] <- -trans*sum(abs(PCw%*%beta))
      SSb[i,j] <- SSb[i-1,j]*(1+Drb[i,j])     # record simulated return
    }
    else{              # open short position, sell target stock will cause transaction cost
      Drb[i,j] <- -trans
      SSb[i,j] <- SSb[i-1,j]*(1+Drb[i,j])     # record simulated return
    }
  }
}
else{                # we have position now
  FLAG2 <- FALSE    # don't want to change PCs when we have position
  signal <- Signal_cal(x[i],gamma_0,i,beta,beta0,Rtn_i)

  if (LONG == TRUE){     # LONG/SHORT Position stop game judgement
    if ((signal < -5)|(signal > 2)|(i-t1 >= 90)){
      FLAG1 <- FALSE    # close out position
      FLAG2 <- TRUE     # need to redo PCA after closing out
    }
  }
  else{
    if ((signal < -2)|(signal > 3)|(i-t1 >= 90)){
      FLAG1 <- FALSE    # close out position
      FLAG2 <- TRUE     # need to redo PCA after closing out
    }
  }

  Drb[i,j] <- Dailyrtn(i,PCw,beta,beta0,x,data_I,sgn)   # for backtest, no transaction cost
  SSb[i,j] <- SSb[i-1,j]*(1+Drb[i,j])     # record simulated return

  if (FLAG1 == FALSE){  # for refinement, with transaction cost
    if (LONG == FALSE){  # close out short position, sell PCs will cause transaction cost
      Drb[i,j] <- Dailyrtn(i,PCw,beta,beta0,x,data_I,sgn) - trans*sum(abs(PCw%*%beta))
      SSb[i,j] <- SSb[i-1,j]*(1+Drb[i,j])     # record simulated return
    }
    else{              # close out long position, sell target stock will cause transaction cost
      Drb[i,j] <- Dailyrtn(i,PCw,beta,beta0,x,data_I,sgn) - trans
```

```
        SSb[i,j] <- SSb[i-1,j]*(1+Drb[i,j])     # record simulated return
       }
      }
    }

    var1b[i,j] <- gamma_0

    # var1r[i,j] <- var
  }
  print(c(j,SSb[enddate,j]))
  plot(SSb[startdate:enddate,j])
}

# Data Present
library(moments)
Eq_Drr <- rowMeans(Drr[startdate:enddate,])
Eq_SSr <- rowMeans(SSr[startdate:enddate,])   # PLOT!
Eq_varr <- rowMeans(var1r[startdate:enddate,]) # PLOT!

min(SSr[enddate,])
max(SSr[enddate,])
annualRtnr <- (SSr[enddate,])^(1/8)-1  # annual return
min(annualRtnr)
max(annualRtnr)
Eq_annualRtnr <- (mean(SSr[enddate,]))^(1/8)-1
Eq_annualRtnr
Eq_AnnualRiskr <- sd(Eq_Drr)*sqrt(250)
Eq_AnnualRiskr
Eq_annualRtnr/Eq_AnnualRiskr

Eq_skewr <- skewness(Eq_Drr)
Eq_skewr
min(skewness(Drr[startdate:enddate,]))
max(skewness(Drr[startdate:enddate,]))

Eq_kurtr <- kurtosis(Eq_Drr)
Eq_kurtr
min(kurtosis(Drr[startdate:enddate,]))
max(kurtosis(Drr[startdate:enddate,]))

Eq_VAR_1d <- -quantile(Eq_Drr,0.05)
Eq_VAR_1d
-quantile(Drr[startdate:enddate,],0.05)

Eq_VAR_1m <- qnorm(0.95)*sd(Eq_Drr)*sqrt(22)-mean(Eq_Drr)
Eq_VAR_1m

# NumPCA[startdate:enddate]  # PLOT!
index_plot<-index(Eq_SSb)
 NUmPCA_plot<-NumPCA[startdate:enddate]
 data_plot_combine<-data.frame(NUmPCA_plot,date_plot,index_plot,Eq_SSb,Eq_varb)
```

```
ggplot(data = data_plot_combine) +
  geom_line(aes(x=date_plot[index_plot], y=Eq_SSb*10, color = "Eq_SSb*10")) +
  geom_line(aes(x=date_plot[index_plot], y=NUmPCA_plot, color = "NUmPCA_plot"))+
  # The value of Eq_SSb is small compare NUmPCA_plot, magnify the variable Eq_SSb 10 times, in order to
clearly display the picture
  ylab("Cum Rtn vs Number of PCs") +
  xlab("Time") +
  theme_bw()
ggsave( file = "backtest Eq_SSb and NUmPCA_plot.png")

Eq_Drb <- rowMeans(Drb[startdate:enddate,])
Eq_SSb <- rowMeans(SSb[startdate:enddate,])   # PLOT!
Eq_varb <- rowMeans(var1b[startdate:enddate,]) # PLOT!
#install.packages("ggplot2")
  library(ggplot2)
  date_plot<-row.names(data)
  date_plot<-as.Date(date_plot)
  date_plot<-date_plot[2000:4000]


#Plot Eq_SSb
index_plot<-index(Eq_SSb)
  data_plot_combine<-data.frame(date_plot,index_plot,Eq_SSb,Eq_varb)
ggplot(data = data_plot_combine) +
  geom_line(aes(x=date_plot[index_plot], y=Eq_SSb, color = Eq_SSb)) +
  ylab("Eq Weight Cum Rtn") +
  xlab("Time") +
  theme_bw()
ggsave( file = "backtest Eq_SSb.png")

#Plot Eq_varb
ggplot(data = data_plot_combine) +
  geom_line(aes(x=date_plot[index_plot], y=Eq_varb, color = Eq_varb)) +
  ylab("Var of residual") +
  xlab("Time") +
  theme_bw()
ggsave( file = "backtest Eq_varb.png")
```