Portfolio management with principal component analysis in a stock market

YU Han

1.Introduction

An investment portfolio is a collection of stocks, bonds, financial derivatives, etc. held by investors or financial institutions. Portfolio management, on the other hand, involves people selecting stocks, bonds, and funds for investment based on their own goals, risk tolerance, and monitoring these investments over time. The main purpose of an investment portfolio is to diversify risk, that is, not to put eggs in one basket. This process can be achieved by selecting unrelated or negatively correlated assets as much as possible. But the real problem is, how to choose from a large number of stocks on the stock market? At this point, principal components can effectively solve this problem. In this project, we will investigate how to achieve passive portfolio management based on the S&P 500 index through principal component analysis.

2.Data Description

In order to track the S&P 500 index more comprehensively, we selected 500 constituent stocks of the S&P 500 as the research objects and collected daily closing price sequences of the S&P 500 index from 2006/01/01 to 2023/12/01 from NYSE and NASDAQ through the API interface of Wind software. Due to our long research period, in order to ensure data integrity to the greatest extent possible, we excluded some data with missing values and conducted research on the remaining 476 stocks.

3.Theory

3.1 Passive portfolio management

Passive portfolio management, also known as index fund management, aims to replicate specific market indices or benchmarks and obtain market (or specific parts of them) returns over time. But passive portfolio management is not equivalent to "setting and forgetting", it also needs to be regularly readjusted to maintain consistency with the selected index. Compared to active portfolio management, passive portfolio management allows for the application of a long-term strategy, which allows us to study portfolio management with principal component analysis in a stock market over a longer period of time.

3.2 Principal Components in Stocks

In Stock market, when choosing stocks by PCA, people usually consider the covariance or correlation matrix of the returns of a set of securities. But there are some problems associated with using a covariance matrix. If there are large differences between the variances of variables, then using a covariance matrix will result in low numbered principal components being dominated by variables that have a large variance. This will impede getting useful information for diversification from a PCA in some cases (Jolliffe, 1986). So here we choose correlation matrix.

The eigenvectors of PCA describe how the returns of stocks are correlated. When the prices are driven by external inputs, the prices of the securities belonging to an eigenvector move together.

3.2 KMO Test

KMO (Kaiser Meyer Olkin) test statistic is an indicator used to compare the simple

correlation coefficient and partial correlation coefficient between variables. The calculation formula is:

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} r_{ij\bullet 1,2\cdots k}^2}$$

The principle is that if there are indeed common factors in the original data, the partial correlation coefficients between the variables should be very small. At this point, the value of KMO is close to 1, which is more suitable for factor analysis. Generally speaking, KMO greater than 0.6 indicates suitability. Although factor analysis and principal component analysis are not complete, the main method in factor analysis is principal component analysis. For stock returns, only when the correlation is relatively high will there be information overlap in the data, which may reduce dimensionality and replace more original variables with fewer principal components. Therefore, in this project, we use KMO to test the shortest length of sliding window that a PCA could be efficiently applied to (Libin Yang, 2015).

3.4 Equal Weight Portfolio Strategy

The Equal Weight Portfolio ,or 1/N investment strategy, means that the strategy to split one's wealth uniformly between the available investment possibilities. In this strategy, the first N principal portfolios are selected, and equal weight is allocated to each principal portfolio. Since eigenvectors remain to be eigenvectors when the signs are flipped, it is important to determine the signs of the eigenvectors $\pm u\_a$ to be adopted. For example, we may fix the sign by requiring the projection of this portfolio to the portfolio with equal weight of stocks to be positive.

4.Result

4.1 Determine window by Window

First we calculate the KMO of the whole dataset is 0.995. Then, to determine the length of window, we choose window length from a year to two years. We find that a window of two year(504trading days) had better KMO statistics, with a lowest of 0.65 and highest of 0.89 during the whole study period. So we decided to apply PCA in rolling window approach with window size two years.

Table.4.1 KMO Test Result of 2-year Window

Date	KMO value
2006/1/1-2007/12/31	0.6489
2008/1/1-2009/12/31	0.8585
2010/1/1-2011/12/31	0.8932
2012/1/1-2013/12/31	0.7262
2014/1/1-2015/12/31	0.7890
2016/1/1-2017/12/31	0.6909
2018/1/1-2019/12/31	0.7283
2020/1/1-2021/12/31	0.8600
2022/1/1-2023/12/31	nan

4.2 Principal Component Analysis on each window

Since we determine the window, now we can make PCA for each period. Take the first period 2006/1/1-2007/12/31 for example. We first calculate the correlation matrix and get result as follow:

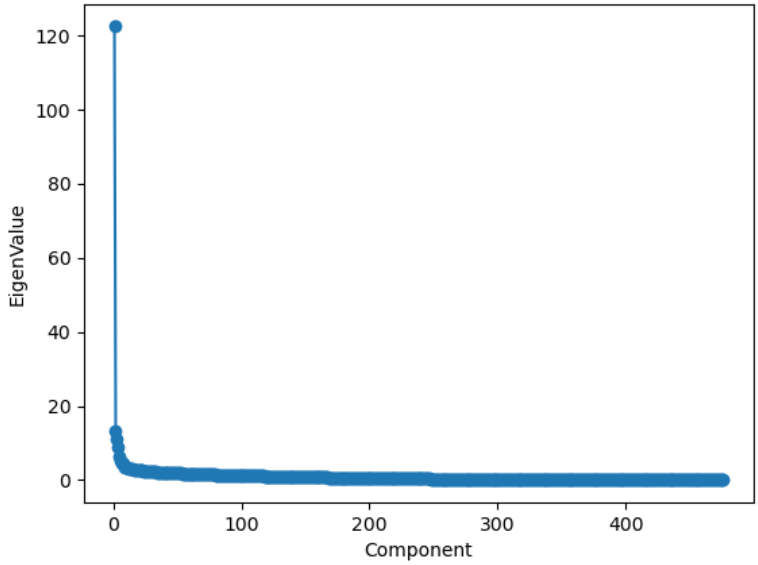


Fig.4.2.1 Eigenvalue of 475 components

Here we can see that most of correlation is between 0.2 and .Then we make PCA on this correlation matrix and get 475 components. It is clearly that the number is too much and we need to delete some components. So How Many Principal Components Should be Included? There are two empirical rules we use:

(1) Cumulative variance: $V_m = \sum_{k=1}^{m} \lambda_k / p$ (p = the sum of all variances = total number of variables). When V_m exceeds 70% or 80%, m is cut off.

(2) Kaiser's rule: Only principal components with eigenvalues > 1 are retained. (If all stocks are uncorrelated, all eigenvalues become 1. So, according to this rule, any principal components with eigenvalues < 1 are not worth retaining.)

Here we followed (Libin Yang, 2015), but changed the deletion criteria. The steps are as follow:

(1) Apply PCA to the correlation matrix of a return dataset.

(2) Associate the variable with the highest coefficient in absolute value with each of

the last m1 principal components that have eigenvalue less than a certain level l.

(3) Repeat the above steps until it satisfies the deletion criteria.

Here we changed the We set the deletion criteria from number to interval, which is around 0.2 and also need to observe the number of components within the deletion criteria simultaneously. We find that after 4 rounds of this kind of deletion, the num of stocks decrease from 475 to 19. What's more, when we do PCA to this 19 stocks correlation matrix, the cumulative variance of the first component increases from 25.72% to 30.07%. According to the cumulative variance rule, here we choose the first 12 components, whose cumulative variances are more than 80%.

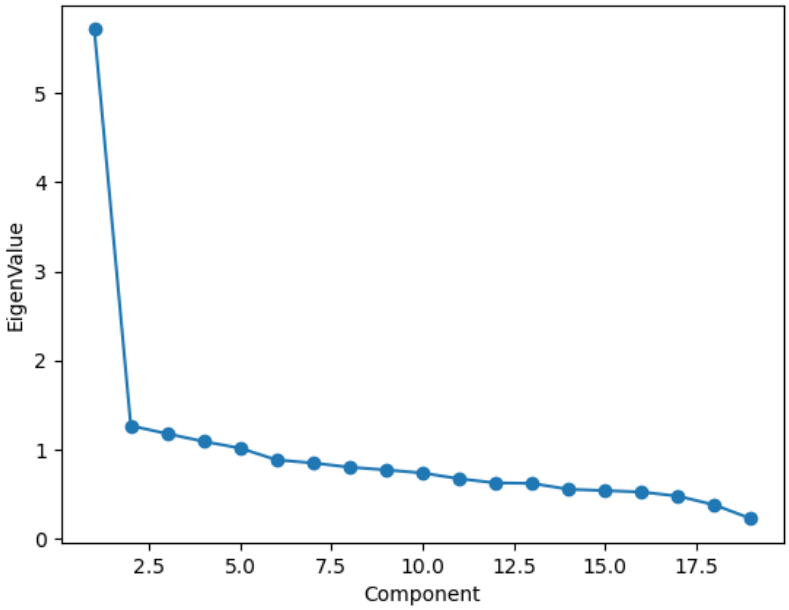


Fig.4.2.2 Eigenvalue of 19 components

We applied this principal component analysis process to other processes, and ultimately obtained the following results:

Table.4.2.1 PCA Result of Different Period

Date	Selection Round	Number of remained stocks	Number of components used
2006/1/1-2007/12/31	4	19	12(82.36%)
2008/1/1-2009/12/31	3	31	13(81.51%)
2010/1/1-2011/12/31	3	29	13(80.54%)
2012/1/1-2013/12/31	3	41	22(80.94%)
2014/1/1-2015/12/31	3	33	17(81.24%)
2016/1/1-2017/12/31	3	42	23(80.28%)
2018/1/1-2019/12/31	3	33	15(80.27%)
2020/1/1-2021/12/31	3	25	9(80.83%)
2022/1/1-2023/12/31	3	35	16(81.23%)

Note: The proportion of cumulative explanatory variance in parentheses

4.3 Build EWP strategy and DRP strategy

Now, we have obtained the PCA and its eigenvalue and eigenvector for each period. So, what strategy should we choose? In class, we learned various strategies. Here, we tried both the Equal Weight Portfolio (or 1/N Strategy) and Diversified Risk Parity strategy, and simulated portfolio management based on these two methods in 4.4. Of course, according to our simulation in 4.4, the DRP strategy is poor, so we will not show the specific weights of the DRP strategy here. This may be explained by the research findings (Georg Ch. Pflug a, Alois Pichler a, David Wozabal 2012) which suggest that in stochastic portfolio decision-making problems with unclear asset returns, a unified

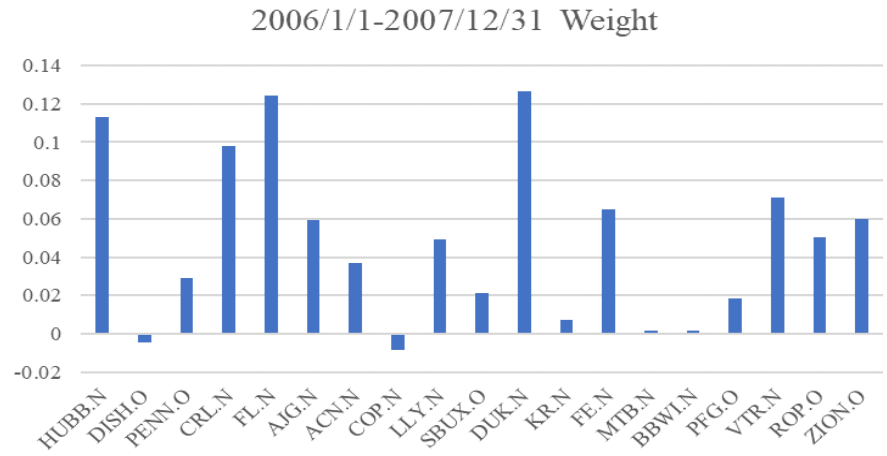
investment strategy or the 1/N rule is a rational strategy。

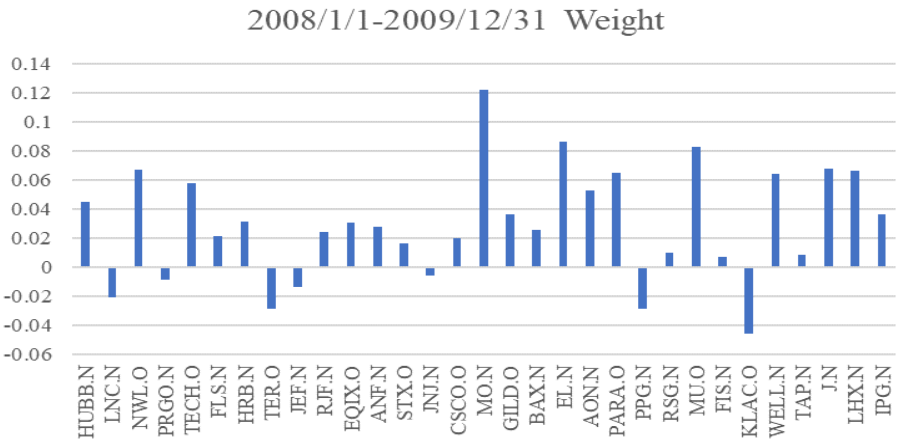
For the EWP strategy, we first need to obtain its projection, whose formula is as follow:

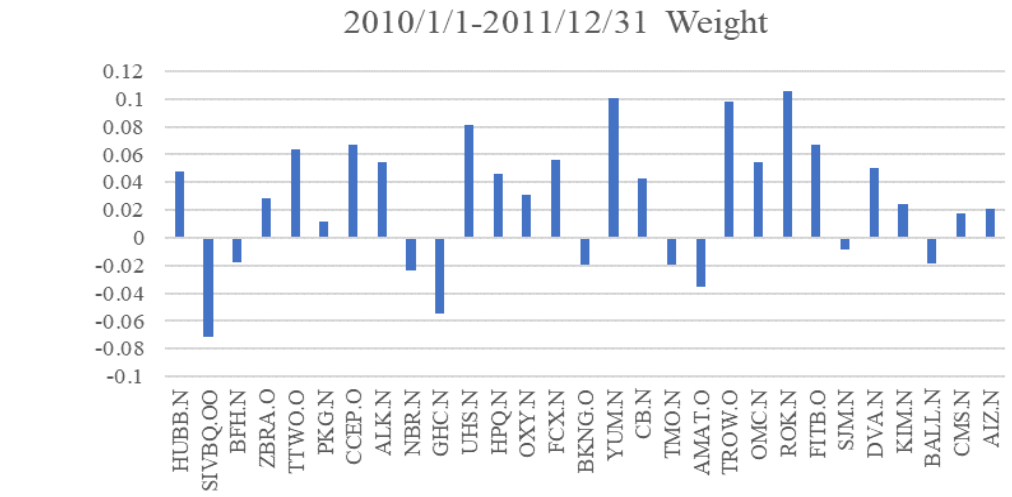
$$\boldsymbol{x}_{PCA} = \bar{\boldsymbol{x}} + \sum_{i=1}^{M} \boldsymbol{u}_i \left[\boldsymbol{u}_i^{\mathrm{T}} (\boldsymbol{x} - \bar{\boldsymbol{x}})\right],$$

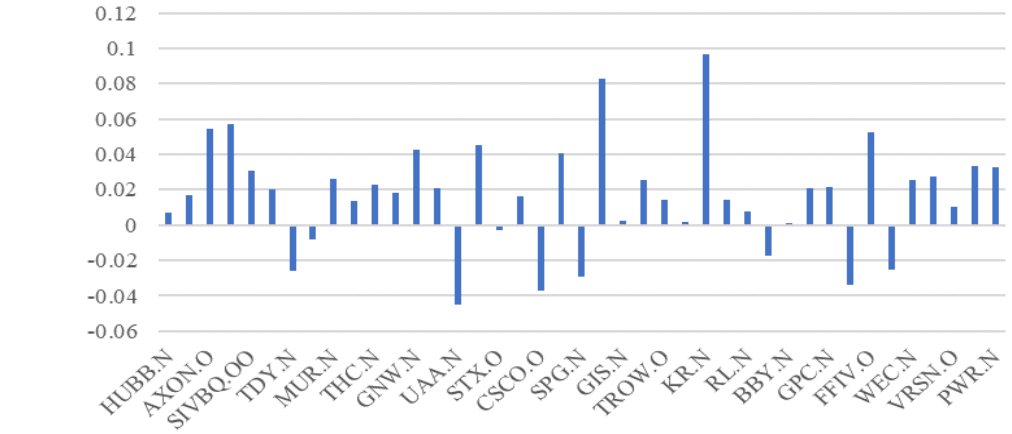
where $\boldsymbol{u}_i$ are the normalized eigenvectors of λ_i.

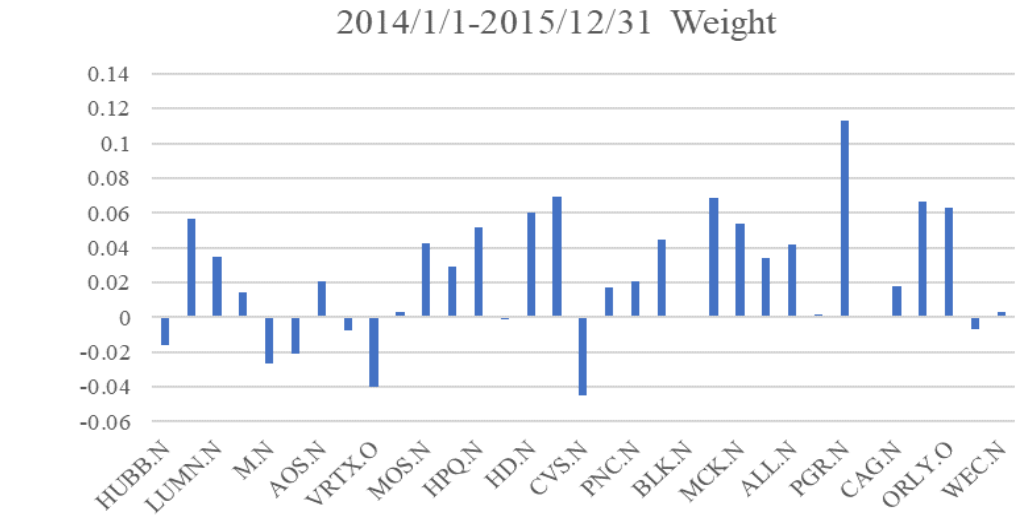
Finally, we get the weight of capital of different stocks as follow, here a positive sign represents a long position, and a negative sign represents a short position:

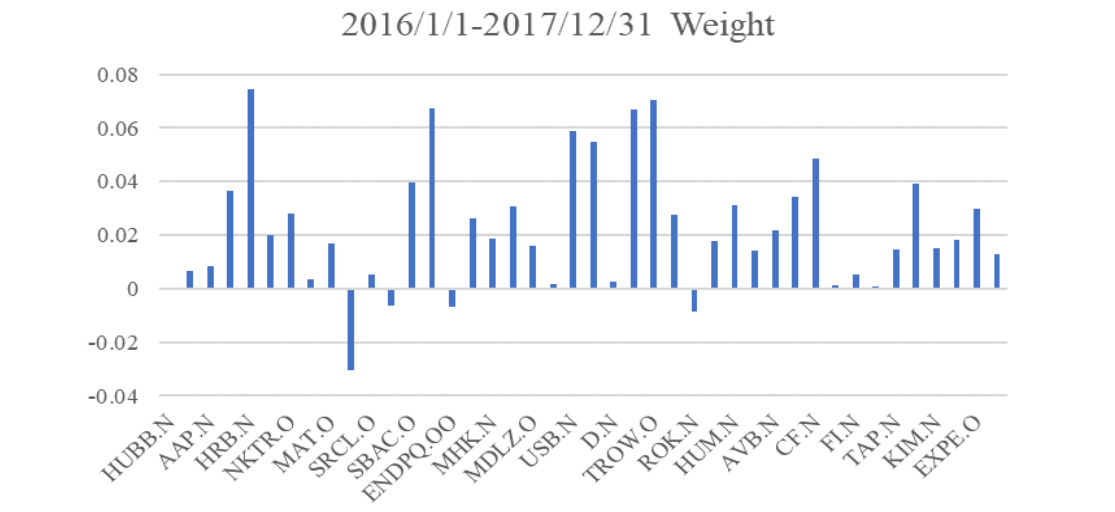


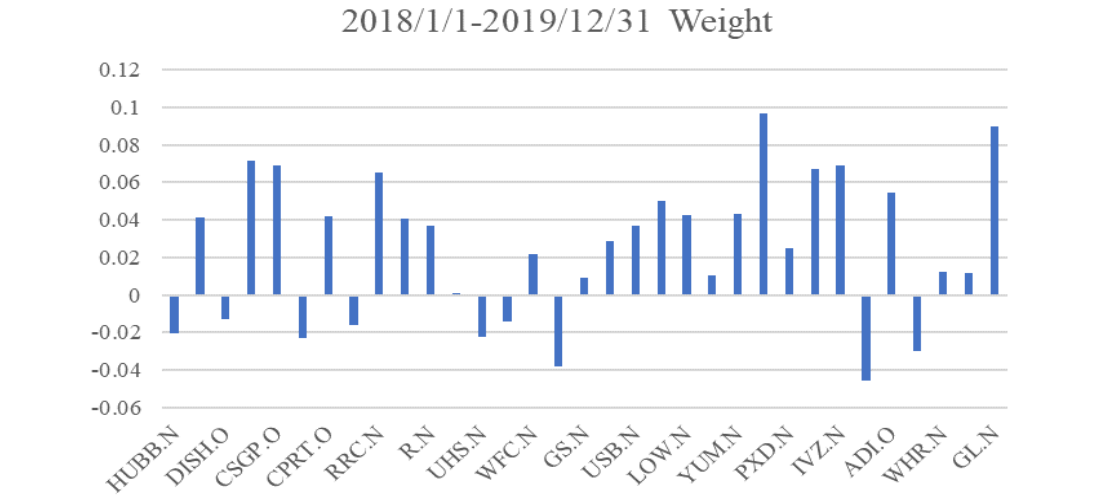


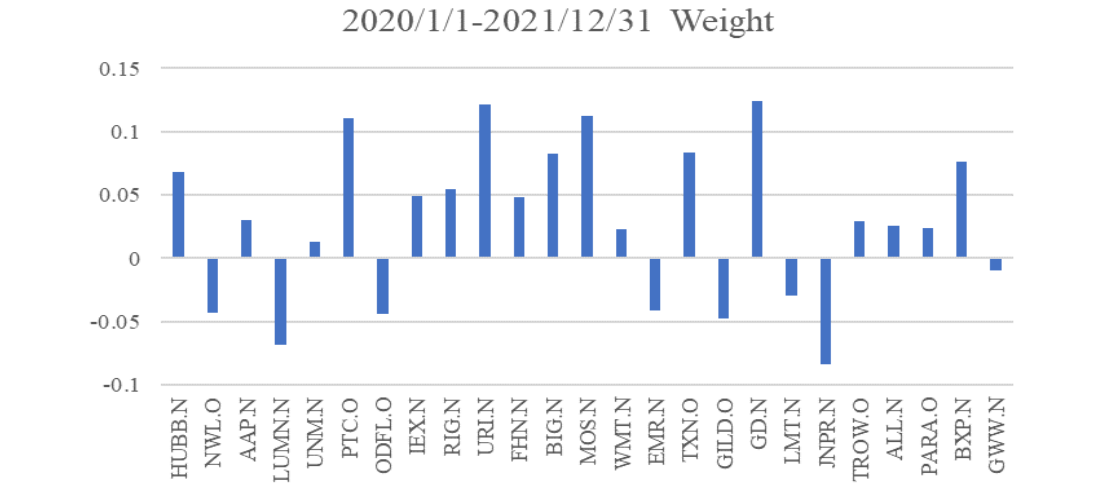

2010/1/1-2011/12/31 Weight
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
HUBB.N
SIVBQ.OO
BFH.N
ZBRA.O
TTWO.O
PKG.N
CCEP.O
ALK.N
NBR.N
GHC.N
UHS.N
HPQ.N
OXY.N
FCX.N
BKNG.O
YUM.N
CB.N
TMO.N
AMAT.O
TROW.O
OMC.N
ROK.N
FITB.O
SJM.N
DVA.N
KIM.N
BALL.N
CMS.N
AIZ.N


2012/1/1-2013/12/31 Weight
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
HUBB.N
AXON.O
SIVBQ.OO
TDY.N
MUR.N
THC.N
GNW.N
UAA.N
STX.O
CSCO.O
SPG.N
GIS.N
TROW.O
KR.N
RL.N
BBY.N
GPC.N
FFIV.O
WEC.N
VRSN.O
PWR.N


2014/1/1-2015/12/31 Weight
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
HUBB.N
LUMN.N
M.N
AOS.N
VRTX.O
MOS.N
HPQ.N
HD.N
CVS.N
PNC.N
BLK.N
MCK.N
ALL.N
PGR.N
CAG.N
ORLY.O
WEC.N


2016/1/1-2017/12/31 Weight
0.08
0.06
0.04
0.02
0
-0.02
-0.04
HUBB.N
AAP.N
HRB.N
NKTR.O
MAT.O
SRCL.O
SBAC.O
ENDPQ.OO
MHK.N
MDLZ.O
USB.N
D.N
TROW.O
ROK.N
HUM.N
AVB.N
CF.N
FI.N
TAP.N
KIM.N
EXPE.O


2018/1/1-2019/12/31 Weight
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
HUBB.N
DISH.O
CSGP.O
CPRT.O
RRC.N
R.N
UHS.N
WFC.N
GS.N
USB.N
LOW.N
YUM.N
PXD.N
IVZ.N
ADI.O
WHR.N
GL.N


2020/1/1-2021/12/31 Weight
0.15
0.1
0.05
0
-0.05
-0.1
HUBB.N
NWL.O
AAP.N
LUMN.N
UNM.N
PTC.O
ODFL.O
IEX.N
RIG.N
URI.N
FHN.N
BIG.N
MOS.N
WMT.N
EMR.N
TXN.O
GILD.O
GD.N
LMT.N
JNPR.N
TROW.O
ALL.N
PARA.O
BXP.N
GWW.N

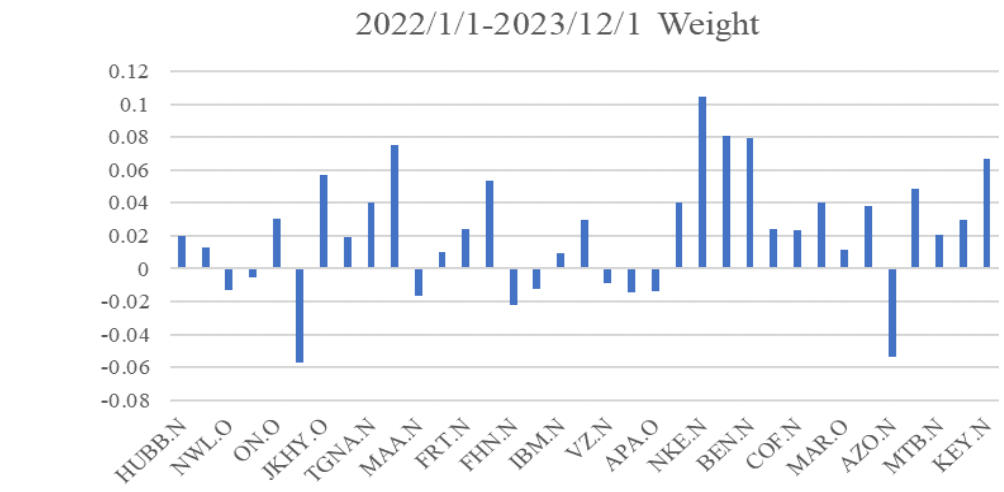


Fig.4.3.1 The Weight of Different Stocks in different Windows

4.4 Portfolio performance

Based on the weights obtained from the EWP strategy and DRP strategy, we construct the investment portfolio. We assume that the investment portfolio allows for short selling, with an initial amount of 1 million, replaced by 1, and no further replenishment or withdrawal of funds will be made unless the position is liquidated. For a window period, the value of the investment portfolio is equal to the closing price of each stock in the portfolio on that day multiplied by the number of shares held, which is determined by the beginning of the period. During the adjustment period between windows, we assume that this adjustment is rigid, meaning that the position adjustment can be fully completed within T+0. We obtained the investment portfolio values constructed by EWP strategy as shown in Figure 4.4.1 and the investment portfolio values constructed by DRP strategy as shown in Figure 4.4.2:

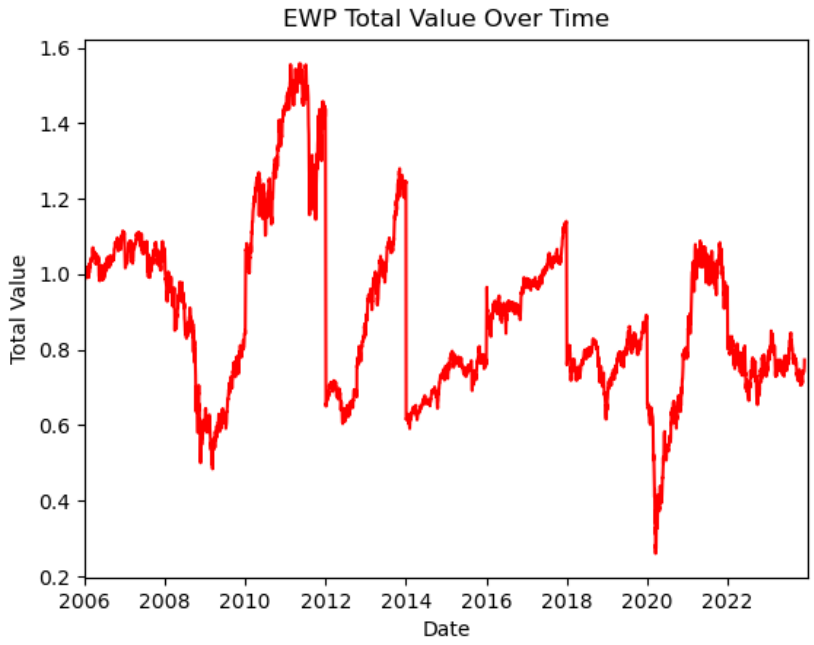


Fig.4.4.1 EWP strategy portfolio

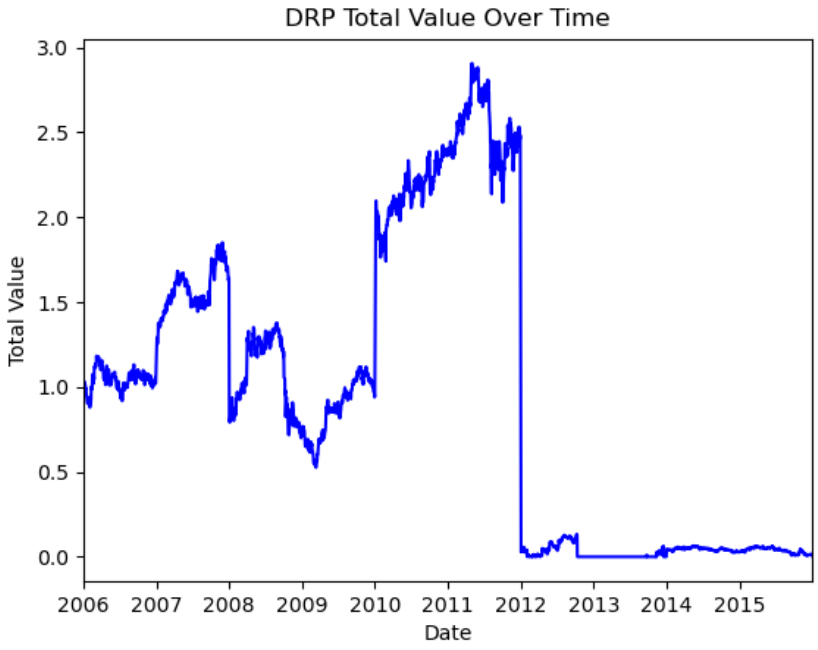


Fig.4.4.1 DRP strategy portfolio

Since the purpose of passive portfolio management is to obtain returns on the S&P 500 index, it is crucial for the portfolio to follow the returns of the S&P 500 index. The return changes between the investment portfolio constructed by EWP strategy and the

S&P 500 index are as Fig.4.4.3.

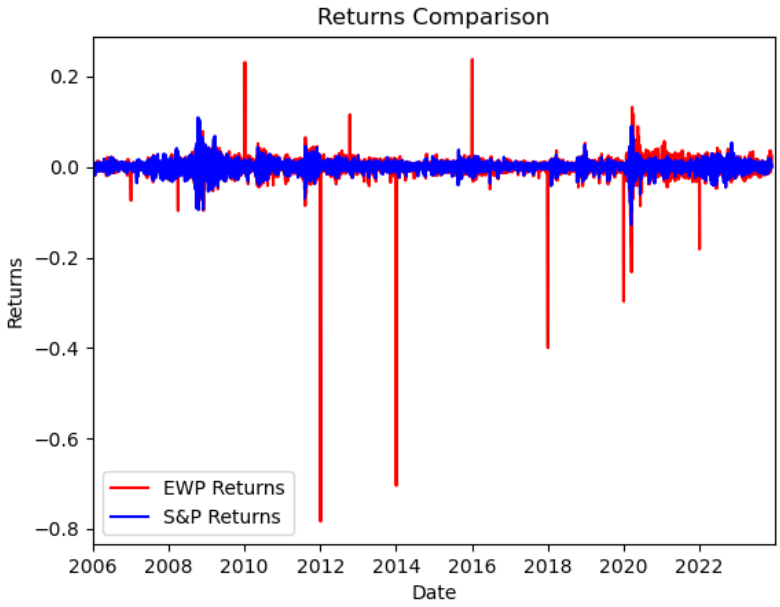


Fig.4.4.3 Returns of EWP portfolio and S&P 500 index

5.Conclusion

(1) Overall, the PCA+EWP method can effectively achieve passive portfolio management based on the S&P 500 index, with only some deviations at certain time points.

(2) Some stock, like HUBB.N(Hubbell) and CSCO.O(Cisco System), play an important role in this portfolio. But the position adjustment before and after the investment portfolio is significant.

(3) Passive portfolio management does not mean that it does not need to be as proactive as active portfolio management. On the contrary, it also requires active adjustment based on indices. For indices with a large number of constituent stocks, such as the S&P 500 index, frequent position adjustments are not a good idea. Two years may be a good idea.

Appendix

A list of 476 stock's code

HUBB.N LNC.N NWL.O AAP.N DISH.O AXON.O LUMN.N FICO.N
SIVBQ.OO SBNY.OO BG.N VNO.N STLD.OACGL.O PCG.N EQT.N
PENN.O PVH.N CSGP.O ON.O CPT.N MOH.N NDSN.O GPS.N LEG.N
FDS.N PRGO.N NOV.N UNM.N BRO.N TECH.O DINO.NCRL.N PTC.O
FLS.N SLG.N XRX.O FTI.N MPWR.O TRMB.O AIV.N POOL.O
KSS.N HRB.N TER.O HOG.N JWN.N BFH.N BIO.N TYL.N TDY.N HP.N
WST.N DXCM.O DPZ.N M.N AMG.N MAC.N ZBRA.O STE.N
ODFL.O WRB.N NKTR.O LVS.N JEF.N NVR.N FL.N IEX.N
MKTX.O MAT.O FLR.N GT.O WAB.N ATO.N TFX.N CE.N SRCL.O
JKHY.O ROL.N CPRT.O AYI.N RRC.N SIG.N PDCO.O TTWO.O
CDNS.O SBAC.O AN.N MUR.N RIG.N MGM.NRMD.N PKG.N AOS.N
R.N ALGN.O EG.N ANSS.O TGNA.N IT.N SWN.N URBN.O
AMD.O RJF.N SNPS.O ENDPQ.OO PBI.N INCY.O IDXX.OOI.N MAA.N
AA.N COO.O MTD.N LNT.O ALB.N CCEP.O AJG.N LKQ.O DLR.N ALK.N
ME.N GPN.N THC.N HOLX.O CNC.N UDR.N CNX.N FRT.N EXR.N
FOSL.O WTW.O CHD.N ILMN.O GNW.N ATI.N JBHT.O WINMQ.OO
O.NNBR.N EQIX.O HSIC.O SWKS.O RCL.N JBL.N GHC.N UHS.N URI.N
X.NMLM.N IGT.N SLM.O UAA.N CLF.N ESS.N TSCO.O VIAV.O ANF.N
MHK.N VRTX.O AME.N FHN.N REGN.O BIG.N FHI.N GRMN.N
ATGE.N PNR.N MNST.O STX.O LRCX.O CCI.N DLTR.OBWA.N
ITT.N MOS.N ACN.N XOM.N AAPL.O CVX.N GE.N MSFT.O IBM.N
GOOGL.O JPM.N WMT.N T.N PG.N ORCL.N WFC.N PFE.N JNJ.N
KO.N BAC.N C.N SLB.N COP.N VZ.N INTC.O PEP.O MRK.N CSCO.O
QCOM.O HPQ.N BRK_B.N OXY.N GS.N DIS.N AMZN.O MCD.N
RTX.N ABT.N UPS.N CAT.N CMCSA.O MMM.N AIG.N HD.N F.N
MDLZ.O AXP.N BA.N MO.N FCX.N USB.N UNH.N APA.O AMGN.O
UNP.N CVS.N HON.O BMY.N HAL.N MET.N EMR.N MDT.N MS.N DE.N
EBAY.OLLY.N TXN.O CL.N DVN.N MRO.N WBA.O BK.N NKE.N
LOW.N TGT.N DHR.N GILD.O PNC.N COST.OSO.N GLW.N SPG.N BAX.N
PRU.N EOG.N FDX.N CSX.O HES.N GD.N BEN.N LMT.N SBUX.O
EXC.O ITW.N BLK.N ELV.N NEM.N KMB.N D.NTRV.N ADP.O BKNG.O
AFL.N CTSH.O NSC.N DUK.N YUM.N SYK.N COF.N NEE.N CCL.N
ADM.N GIS.N STT.N SCHW.N CB.N JNPR.N CMI.N TMO.N JCI.N
AMT.N AMAT.O CME.O MCK.N K.NPCAR.O APD.N TJX.N TFC.N
ETN.N EL.N PSA.N NOC.N WMB.NL.N AON.N WM.N CRM.N BDX.N
AEP.O NTAP.O TROW.O BIIB.O ADBE.O VLO.N ALL.N EQR.N
WYNN.O INTU.O MMC.NSYY.N PEG.N TT.N TPR.N PARA.O A.N
PH.N AMP.N PPG.N MSI.N ED.N CAH.N KR.N NUE.N GEN.O PGR.N
PEAK.N OMC.N ROK.N BXP.N WY.N ISRG.O IP.N MAR.O SRE.N
SWK.N NFLX.O DOV.N PPL.N CHRW.O HSY.N HIG.N RL.N ETR.N
PXD.N CI.N EIX.N SPGI.N ECL.N IVZ.N HST.O HUM.N ZBH.N ADI.O

XEL.O AZO.N RSG.N PAYX.O FE.N BBY.N MU.O BSX.N FITB.O
COR.N EXPD.N MTB.N NVDA.O VFC.N CPB.N BBWI.N AVB.N
CAG.N AES.N PFG.O FIS.N ADSK.O CF.N EW.N DGX.N GWW.N
CLX.N FAST.O LUV.N LH.N APH.N FI.N RF.N SHW.N CBRE.N
WDC.O VTR.N ICE.N GPC.N ROP.O ROST.OSJM.N DTE.N DVA.N
ORLY.OWAT.N MCO.N KEY.N KLAC.O WELL.N TAP.N TXT.N FFIV.O
CNP.N HRL.N KMX.N TSN.N KIM.N OKE.N WEC.N MCHP.O EMN.N
AKAM.O AEE.N DRI.N EA.O WHR.N CMA.N J.N HAS.O LHX.N MKC.N
VRSN.O IRM.N IPG.N ES.N FMC.N PLD.N EXPE.OBF_B.N BALL.N
VMC.N HBAN.O CTRA.N NI.N CINF.O NRG.N XRAY.O GL.N
MAS.N IFF.N NDAQ.O CMS.N EFX.N PWR.N PNW.N RHI.N AVY.N
CTAS.OZION.O SEE.N AIZ.N STZ.N DHI.N SNA.N RVTY.NPHM.N LEN.N

Reference

1. Jolliffe, I. T. (1986). Principal Component Analysis. Springer, New York

2. Yang, L. (2015). An Application of Principal Component Analysis to Stock Portfolio Management.

3. Georg Ch. Pflug, Alois Pichler, David Wozabal (2012). The 1/N investment strategy is optimal under high model ambiguity, Journal of Banking & Finance,