Portfolio Construction Using Principle Component Analysis

By
Huanting Chen

A Master Project
Submitted to the Faculty
of the
Worcester Polytechnic Institute
in partial fulfillment of the requirements of the
Professional Degree of Master of Science

In

Financial Mathematics

August 2014

Approved:

Dr. Marcel Blais, Advisor

Dr. Luca Capogna, Head of Department

Table of Contents

Abstract

Principal Components Analysis (PCA) is an important mathematical technique widely used in the world of quantitative finance. The ultimate goal of this paper is to construct a portfolio with hedging positions, which is able to outperform the SPY benchmark in terms of the Sharpe ratio. Mathematical techniques implemented in this paper besides principle component analysis are the Sharpe ratio, ARMA, ARCH, GARCH, ACF, and Markowitz methodology. Information about these mathematical techniques is listed in the introduction section.

Through conducting in sample analysis, out sample analysis, and back testing, it is demonstrated that the quantitative approach adopted in this paper, such as principle component analysis, can be used to find the major driving factor causing movements of a portfolio, and we can perform a more effective portfolio analysis by using principle component analysis to reduce the dimensions of a financial model.

Introduction

Sharpe Ratio

The Sharpe ratio is used to measure how much of a portfolio's returns are caused by a smart investment decision or a result of excess risk. The Sharpe ratio is a risk-adjusted measure of return that is often used to evaluate the performance of a portfolio. A larger Sharpe ratio indicates that the portfolio has a better ability to generate a higher return on a risk adjust basis compared to a portfolio with lower Sharpe ratio.

Most measurements of performance are computed with historical data; however, depending on different predicted relationships these measurements can be justified. Ex-post results are used in practical implementations; however, theoretical discussions focus on ex-ante values. Implicitly or explicitly, it is assumed that historical results have at least some predictive ability, as practical implementations use ex-post results while theoretical discussions focus on ex-ante values. (Sharpe 1)

The term ex ante (sometimes written ex-ante) is a phrase meaning "before the event," so the ex-ante Sharp Ratio is calculated using the expected returns. The ex-post Sharpe Ratio is calculated using realized returns instead of expected returns.

The Ex-Ante Sharpe Ratio

According to Sharpe we give the dentition of ex-ante Shapre ratio here. Let R_f represent the return on fund F in the forthcoming period and R_B the return on a benchmark portfolio or security. Define d, the differential return, as

(1) $\tilde{d}=\widetilde{R_F}-\tilde{R}_B$.

In the equations, the tildes over the variables indicate that the exact values may not be known in advance.

Let $\bar{d}$ be the expected value of d and σ_d be the predicted standard deviation of d. The ex ante Sharpe Ratio (S) is

(2) $s\equiv \frac{\bar{d}}{\sigma_d}$.

In this version the ratio indicates the expected differential return per unit of risk associated with the differential return. (Sharp 3)

The Ex-Post Sharpe Ratio

According to Sharpe the ex-post Sharpe Ratio is defined by

$$(3)\ \bar{D} \equiv \frac{1}{T}\sum_{t=1}^{T} D_t,$$

where R_{Ft} is the return on the fund in period t, R_{Bt} is the return on the benchmark portfolio or security in period t, and D_t is the differential return in period t.

Let $\bar{D}$ be the average value of D_t over the historical period from t=1 through t=T, and let σ_D be the sample standard deviation over the same historical period,

$$(4)\sigma_D \equiv \sqrt{\frac{\sum_{t=1}^{T}(D_t-\bar{D})^2}{T-1}}.$$

The Ex-post, or historic Sharpe Ratio, is given by

$$(6) S_h \equiv \frac{\bar{D}}{\sigma_D}.$$

(Sharp 3)

Log Return

There are several measurements which are used to assess the return on an asset, such as the net return, the gross return, and the log return. The log return is used to calculate the daily price returns as log (p2/p1). Here p2 is the price of a stock on the second day, p1 is the price on the first day.

Log returns are denoted by r_t and defined as

$$r_t = \log(1+R_t) = \log\left(\frac{P_t}{P_{t-1}}\right) = p_t - p_{t-1},$$

where $p_t = \log(P_t)$ is called the log price. Log returns are approximately equal to returns because if x is small, then

$\log(1 + x) \approx x$, s.t $|x| < 0.1$.

The log return is widely used in financial world because the use of log returns simplifies multi-period returns. Instead of the product as in the case of gross returns, a k-period log return is simply the sum of the single-period log returns. To see this note that the k-period log return is

$$r_t(k) = \log\{1 + R_t(k)\} = \log\{(1+R_t)\cdots(1+R_{t-k+1})\} =$$

$$\log(1+R_t)+\cdots+\log(1+R_{t-k+1}) = r_t + r_{t-1} + \cdots + r_{t-k+1}.$$ (Ruppert 26)

Principle component analysis

Principle component analysis (PCA) is a widely used dimension reduction technique. Through finding structure in the covariance or correlation matrix we use this structure to locate low-dimensional subspaces containing most of the variation in the data.

PCA starts with a sample $Y_i = (Y_{i,1},\ldots,Y_{i,d})$, $i = 1,\ldots,n$, of d-dimensional random vectors with mean vector μ and covariance matrix Σ. One goal of PCA is finding "structure" in Σ. (Ruppert 443)

According to Haugh, in the context of risk management, we take this vector to represent the (normalized) changes, over some appropriately chosen time horizon, of an n-dimensional vector of risk factors. These risk factors could represent security price returns, returns on futures contracts of varying maturities, or changes in spot interest rates, again of varying maturities.

Let $Y = (Y1, \ldots, Yn)^T$ denote an n-dimensional random vector with variance-covariance matrix, Σ. The goal of PCA is to construct linear combinations in such a way that:

Let $Y = (Y1, \ldots, Yn)^T$ denote an n-dimensional random vector with variance-

covariance matrix, Σ. The goal of PCA is to construct linear combinations

$$P_i = \sum_{j=1}^{n} w_{ij} Y_j, \text{ for } i=1,\ldots,n$$

in such a way that:

(1) The P_i's are orthogonal so that $E[P_i P_j] = 0$ for $i \neq j$, and

(2) The P_i's are ordered so that: (i) P_1 explains the largest percentage of the total variability in the system and (ii) each P_i explains the largest percentage of the total variability in the system that has not already been explained by $P_1, \ldots, P_{i-1}$. (Haugh 7)

According to Hauge, if the normalized random variables satisfy $E[Y_i] = 0$ and $Var(Y_i) = 1$ it is very common to apply PCA in practice. This is achieved by subtracting the means from the original random variables and dividing by their respective standard deviations. We do this to ensure that no single component of Y can influence the analysis by virtue of that component's measurement units. We will therefore assume that the Y_i's have already been normalized.

The key tool of PCA is the spectral decomposition from linear algebra which states that any symmetric matrix $A \in \mathbb{R}^{n \times n}$ can be written

$$A = \Gamma \Delta \Gamma^T,$$

where (i) Δ is a diagonal matrix, diag(λ 1, ⋯, λ n), of the eigenvalues of A

which, without loss of generality, are ordered so that

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and (ii) Γ is an orthogonal matrix with the i^{th} column of Γ containing the i^{th} standardized eigen-vector, γ_i of A. The orthogonality of Γ implies $\Gamma \Gamma^T = \Gamma^T \Gamma = I_n$.

Since Σ is symmetric we can take $A = \Sigma$ in (1), and the positive semi-definiteness of Σ implies $\lambda_i \geq 0$ for all $i = 1,...,n$. The principal components of Y are then given by $P = (P_1,...,P_n)$ satisfying $P = \Gamma^T Y$. (Haugh 8)

GARCH Model

The generalized autoregressive conditional heteroskedasticity (GARCH) model is commonly used for financial time series due to the importance of variance in calculating derivative prices and modeling risk. (DeWeese 5) In some financial models it is assumed that returns are stationary, as it is in the GARCH model, though this may not adequately capture extreme variation in returns. To alleviate this assumption this paper uses the Student's t-distribution in the GARCH model.

The GARCH model is used to foresee the future conditional variance of a time series.

The GARCH (1,1) model (with a Student t-distribution),

$$\sigma_{t+1}^2 = \omega + \alpha X_t^2 + \beta\sigma_t^2 \quad s.t. X_t = \sigma_t\epsilon_t \quad s.t. \epsilon_t \sim St(v),$$

$$\mathbb{E}(\epsilon) = 0, and \mathbb{E}(\epsilon\epsilon^T) = \frac{v}{v-2}\mathbf{I},$$

ϵ_t is a white noise process, *St(v)* is the Student t-distribution with degrees of freedom, and **I** is the identity matrix, $\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$ (DeWeese 5)

ARMA Model

The definition of ARMA model used in this paper is defined by Ruppert. An ARMA(p,q) model combines both AR and MA terms and is defined by the equation

$$(Yt - \mu) = \phi\,1(Yt-1-\mu)+\cdots+\phi\,p(Yt-p-\mu)+\varepsilon\,t+\theta\,1\varepsilon_t-1+\cdots+\theta\,q\varepsilon_t\text{-}\,q,$$

(1)

which shows how Yt depends on lagged values of itself and lagged values of the white noise process. Equation (1) can be written more succinctly with the backwards operator as

$$(1 - \phi\,1\,B - \cdots - \phi\,p\,Bp)(Yt - \mu) = (1 + \theta_1\,B + \cdots + \theta\,q\,Bq)\varepsilon_t.$$

(2)

A white noise process is ARMA(0,0) since if $p = q = 0$, then (2) reduces to

$$(Yt - \mu) = \varepsilon_t. \quad (225)$$

Autocovariance function (ACF)

The definition of the autocovariance function is also from the Ruppert text. The autocovariance function is used to find how far back in time the data (ex. Data at t, t-1, t-2,..., t-n) is effective for use in predicting the future movements of the stocks.

Assume we observe Y1, . . . , Yn from a stationary process. We can use the sample mean Y and sample variance S^2 to estimate the mean μ and variance σ^2 of the process. (Ruppert 225)

To estimate the autocovariance function, we use the sample autocovariance function

$$\hat{\gamma}(h) = n^{-1} \sum_{j=1}^{n-h} (Y_{j+h} - \bar{Y})(Y_j - \bar{Y}) . \qquad (3)$$

Equation(3) is an example of the usefulness of parsimony induced by the stationarity assumption. (Ruppert 206)

ARCH (p) Models

According to Ruppert, if we let ε_t be Gaussian white noise with unit variance, then the ARCH (q) process is defined as

$$\alpha_t = \sigma_t \varepsilon_t,$$

where

$$\sigma_t = \sqrt{\omega + \sum_{i=1}^{p} a_i a_{t-i}^2}.$$

According to Sharpe, similar to an ARCH (1) process, an ARCH (q) has a constant mean (both conditional and unconditional) and a constant unconditional variance, but its conditional variance is nonconstant. In fact, the ACF of α_{2t} is the same as the ACF of an ARCH (q) process. (Ruppert 482)

Markowitz's Mean-Variance Analysis

Markowitz's Mean-Variance Analysis is used to gain a set of optimized hedging positions in order to minimize the variance of the return. The same kind of analysis can be used when there exists a riskless asset.

The budget constraint is:

$$w' e + w_0 = 1 \Longleftrightarrow w_0 = 1 - we,$$

where w is the vector of weights of the n risky assets, and R is the vector of returns. Let Rf denote the return of the riskless asset. Let $w0$ denote the percentage of wealth invested on this riskless asset , and e is the vector $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

The new optimization program is:

$$min\ wVw,$$

$$\text{with } w'\bar{R} + (1 - w'e)R_f = \mathrm{E}[R_p],$$

The Lagrangian function associated to the problem above is

$$\mathrm{L}(\mathrm{w}, \lambda) = \mathrm{w}' \mathrm{Vw} + \lambda \left(\mathrm{E}[R_P] - w'\bar{R} - \left(1 - w'e\right) R_f \right),$$

Thus we have to solve:

$$min\ \mathrm{L}(\mathrm{w}, \lambda).$$

The first-order conditions, which are also necessary and sufficient, are given by:

$$\frac{\partial L(w,\lambda)}{\partial w} = 2Vw - \lambda(\bar{R} - eR_f) = 0,$$

$$\frac{\partial L(w,\lambda)}{\partial \lambda} = \mathrm{E}[\mathrm{RP}] - \mathrm{w}'\bar{R} - (1 - \mathrm{w}'\ \mathrm{e})\ \mathrm{Rf} = 0.$$

Then the optimal portfolio at the level E [RP] satisfies:

$$\mathrm{W}=V^{-1}(\bar{R}-eR_f)\frac{\mathrm{E}[\mathrm{R_P}]-R_f}{(\bar{R}-eR_f)'V^{-1}(\bar{R}-eR_f)},$$

Its variance is given by

$$\sigma^2(R_p)=w'Vw=\frac{(\mathrm{E}[R_p]-R_p)^2}{J},$$

where $J=B-2ARf+CR_f^2$ is non-negative.

The classical mean-variance setting of Markowitz assumes a one-period model in which investors are only concerned with the mean and variance of their portfolio returns. All other things being equal, such investors prefer portfolios with greater mean returns and smaller return variances. (Prigent, 75)

Objective

As mentioned previously, our main objective is to construct a stock portfolio that outperforms the S&P 500 index using SPY as the proxy since the S&P 500 index is not a tradable instrument.

The SPDR Trust Series I (NYSE: SPY) is an exchange traded fund (ETF) that tracks the performance of the S&P 500 index. The fund is managed by the State Street Global Advisors and is traded on the New York Stock Exchange. Shares in the fund can be bought through any licensed Series 6 stock broker.

Each unit (share) of the fund is priced to reflect 1/10 of the value of the S&P 500 index. The shares trade continuously on the market. They pay the aggregate dividend of the companies in the S&P 500 index.

The plan to achieve this objective is by principal component analysis (PCA) and Markowitz portfolio theory. The reasons behind this are twofold. First, PCA will allow us to find the stocks that can best represent the market, whereby we construct our market portfolio. Second, the tangency portfolio has the highest Sharpe ratio. If we construct our portfolio according to the tangency portfolio weights solved by:

$$\sum_{j=1}^{10} \sigma_{i,j} w_j - \lambda_1 \mu_i - \lambda_2 = 0, i = 1, \dots, 10$$

$$\sum_{j=1}^{10} w_j \mu_j = \mu$$

$$\sum_{j=1}^{10} w_j = 1$$

then our portfolio will have the highest Sharpe ratio possible.

With these 2 intuitions, we plan to classify the S&P 500 stocks into 10 sectors. In each sector we will construct a sector index using the principal components of the sector stocks. We will then find the tangency portfolio weights of these 10 sector indices.

The experiment will first be done in-sample to get a sense of the performance retrospectively. We will then perform out-of-sample test to evaluate the real performance. The performance metric we use is the Sharpe ratio.

Data

Data sources were collected from Yahoo finance by downloading the daily closing prices for the stocks in the file by the following codes:

```
df<-
read.csv("http://ichart.finance.yahoo.com/table.csv?s=IBM&ignore=.csv",
stringsAsFactors=F)
```

The period we use is roughly 10 years in length, from April 2003 to April 2013. We then calculated the daily price returns as log(p2/p1). These daily price returns time series are the subjects of our statistical analysis.

Analysis

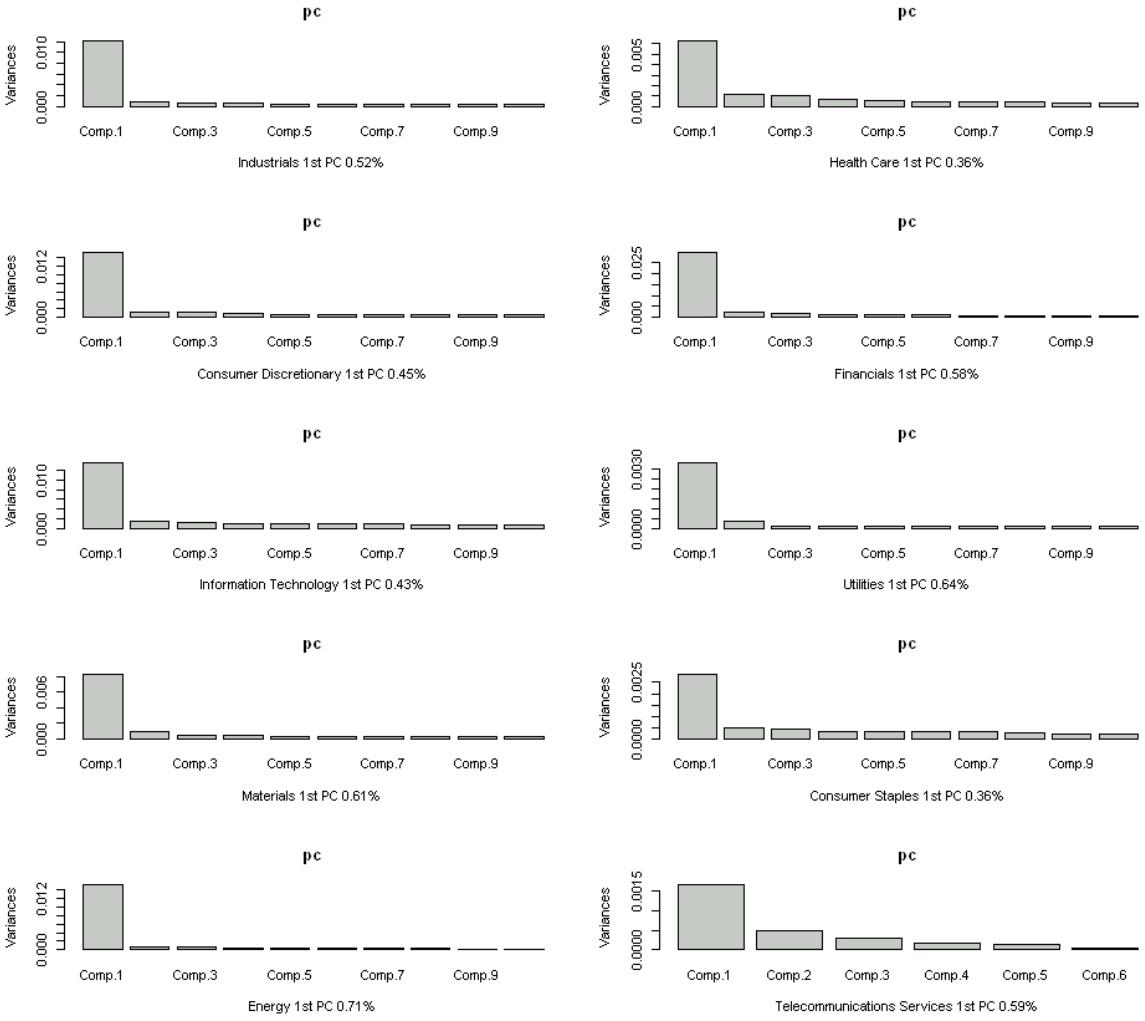
In-sample analysis

Figure 1 in the attachments shows the screen plots of the 10 PCA's we conducted on the 10 sectors of the S&P 500 stocks. As indicated by the graphs, all 10 sectors' first principal components capture the majority of the variances of the sectors. Many of them are over 50%. The minimum is 36%.

These results indicate that the returns data of all 10 sectors are well structured. With only the first principal component we might be able to represent the sector fairly well. We then proceeded to construct 10 sector indices with the only the first principal components,

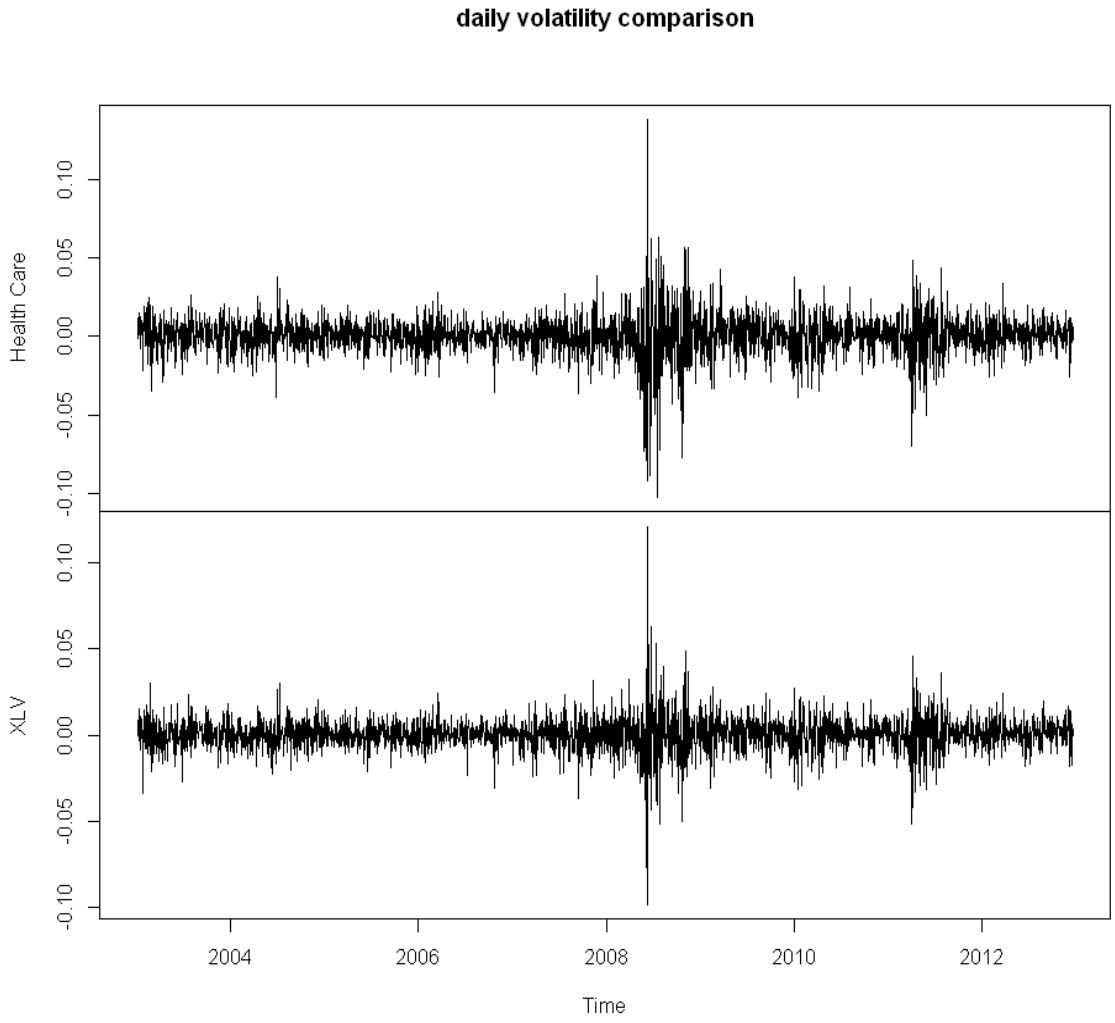
$$W_1 = \text{argMAX Var}\{w^T X\}.$$

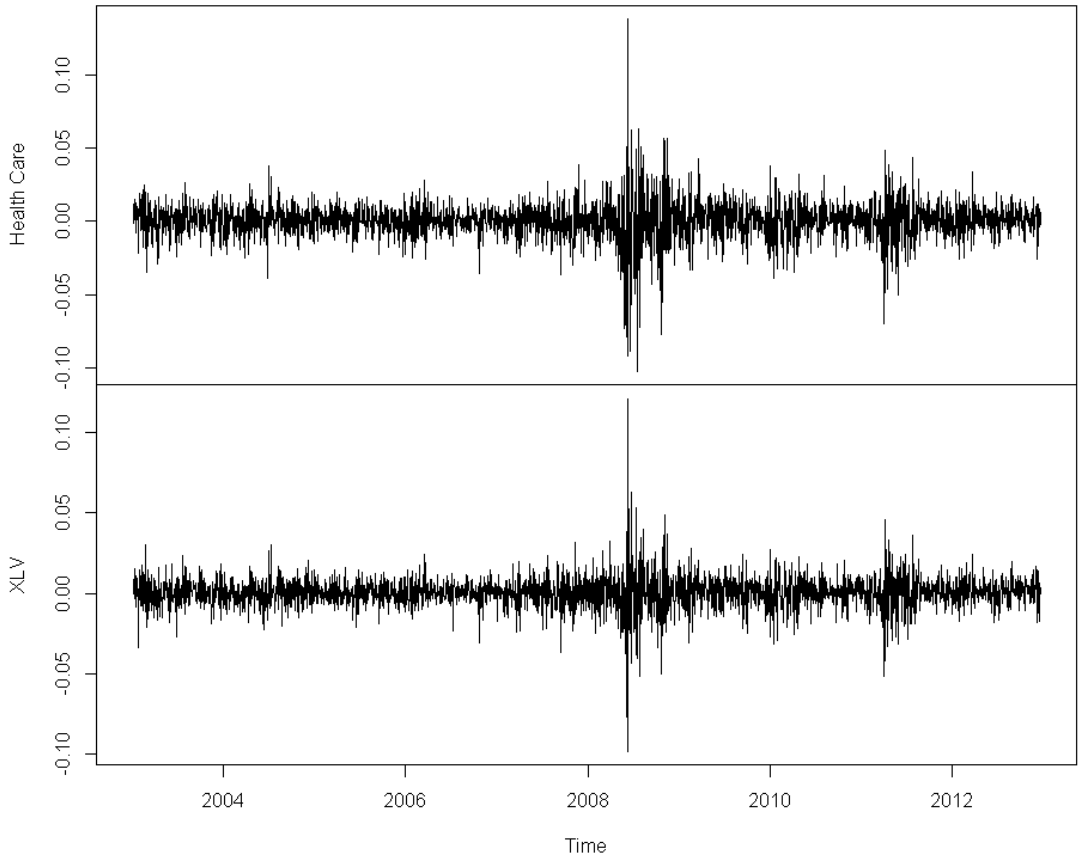
Figure 1: In sample Apr 2003 – Apr 2013, PCA

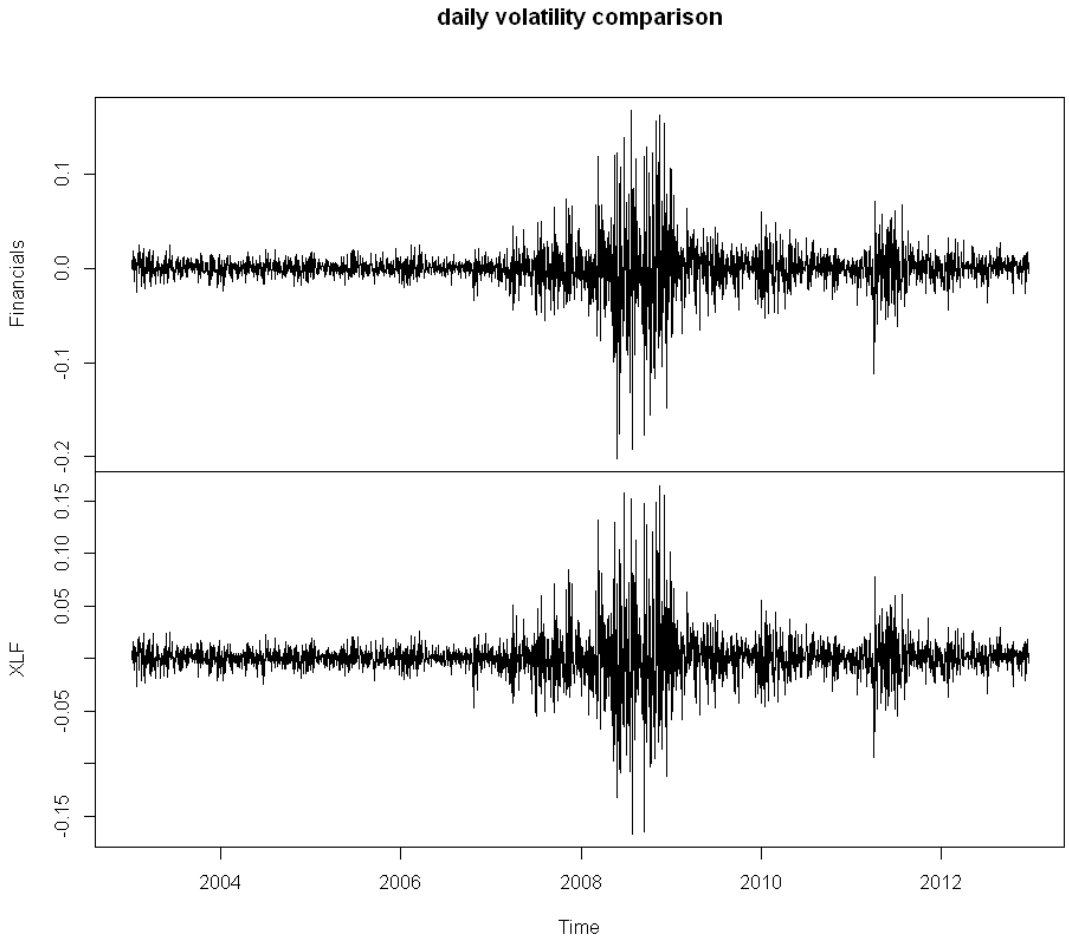


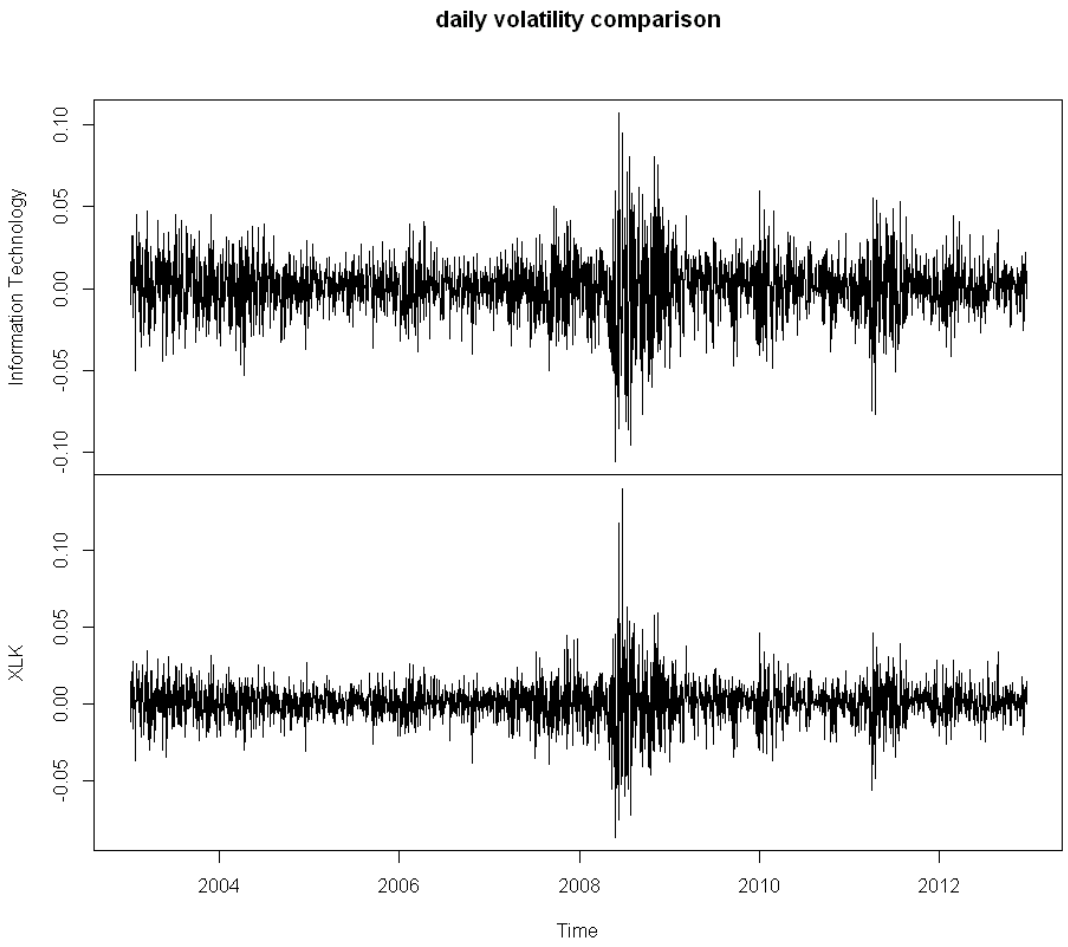
To verify the 10 sector indices we constructed, we compared the 10 sector daily return time series with those of the corresponding sector ETFs. In Figure 2 we can see that the characteristics are similar between each pair. They all have similar volatility clusters in the same periods.

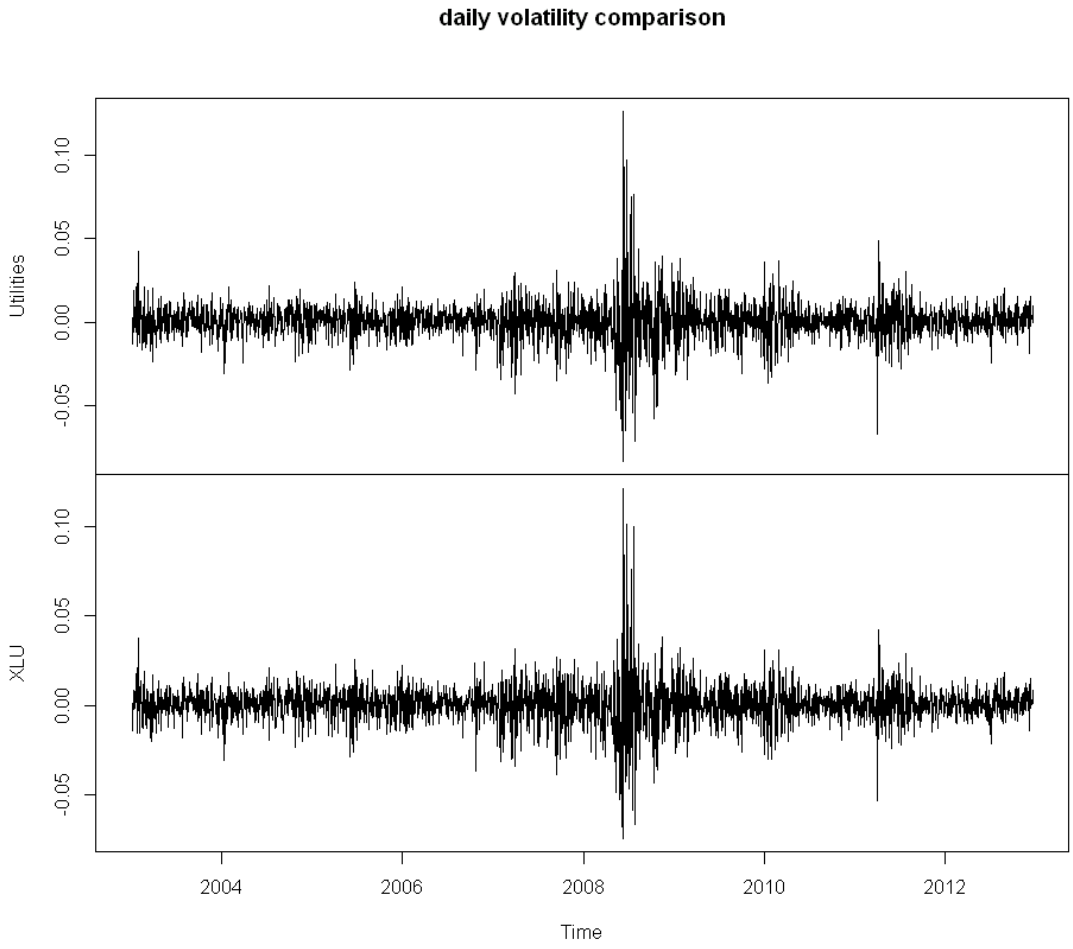
Figure 2: daily returns of sector indices vs sector ETFs

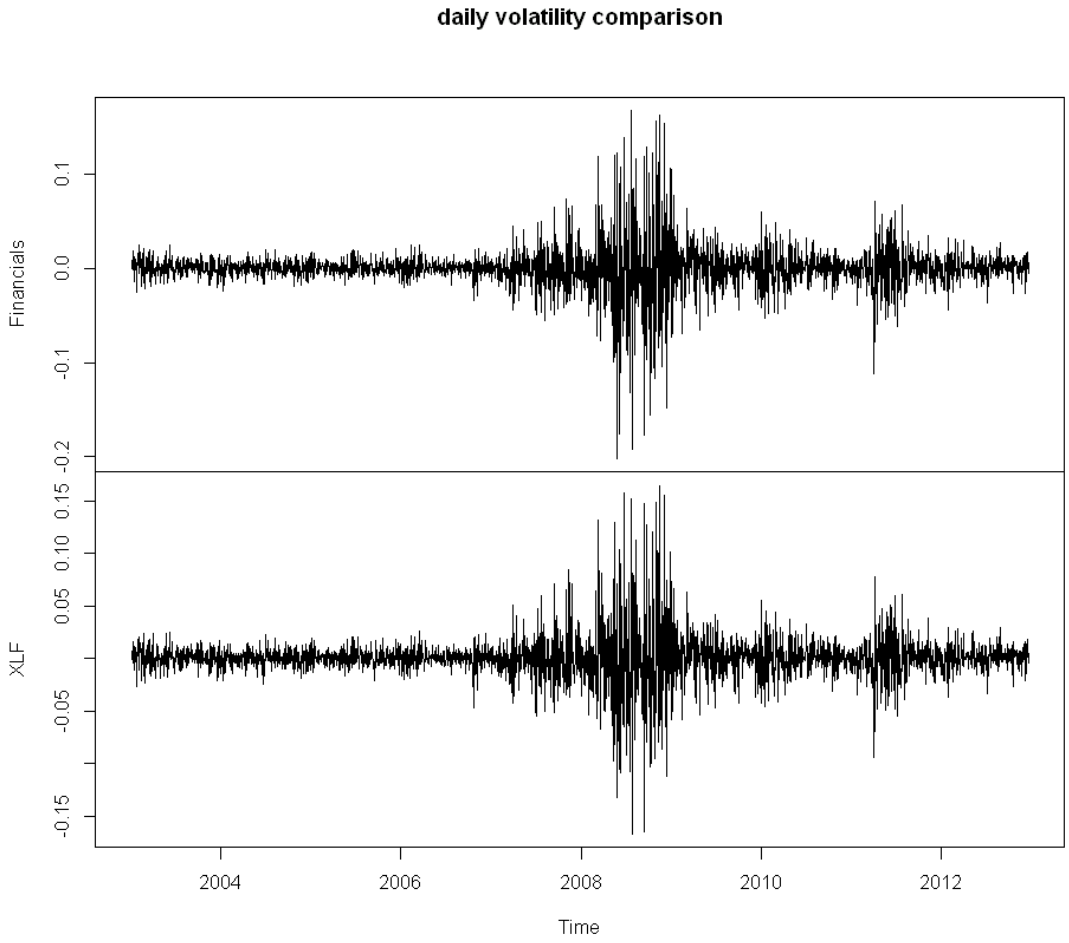


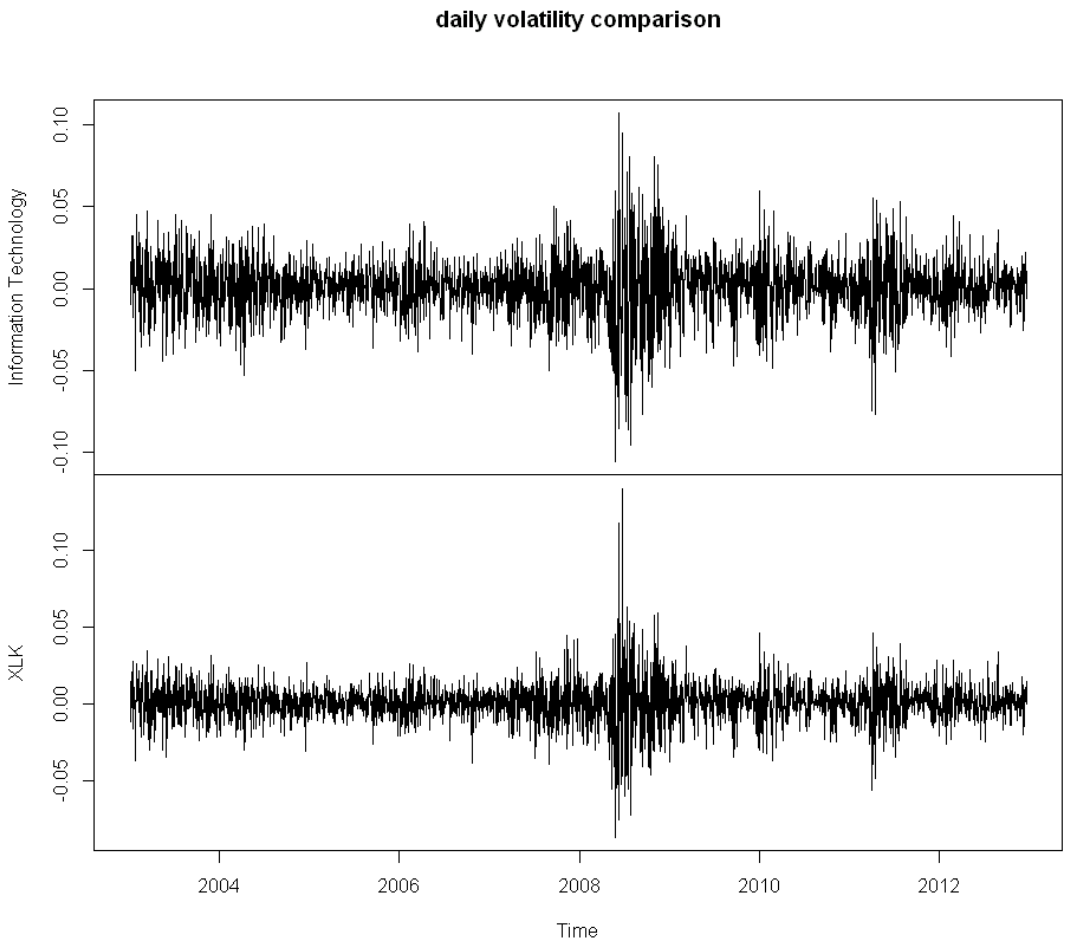

daily volatility comparison
Health Care
XLV
0.10
0.05
0.00
-0.05
-0.10
2004
2006
2008
2010
2012
Time

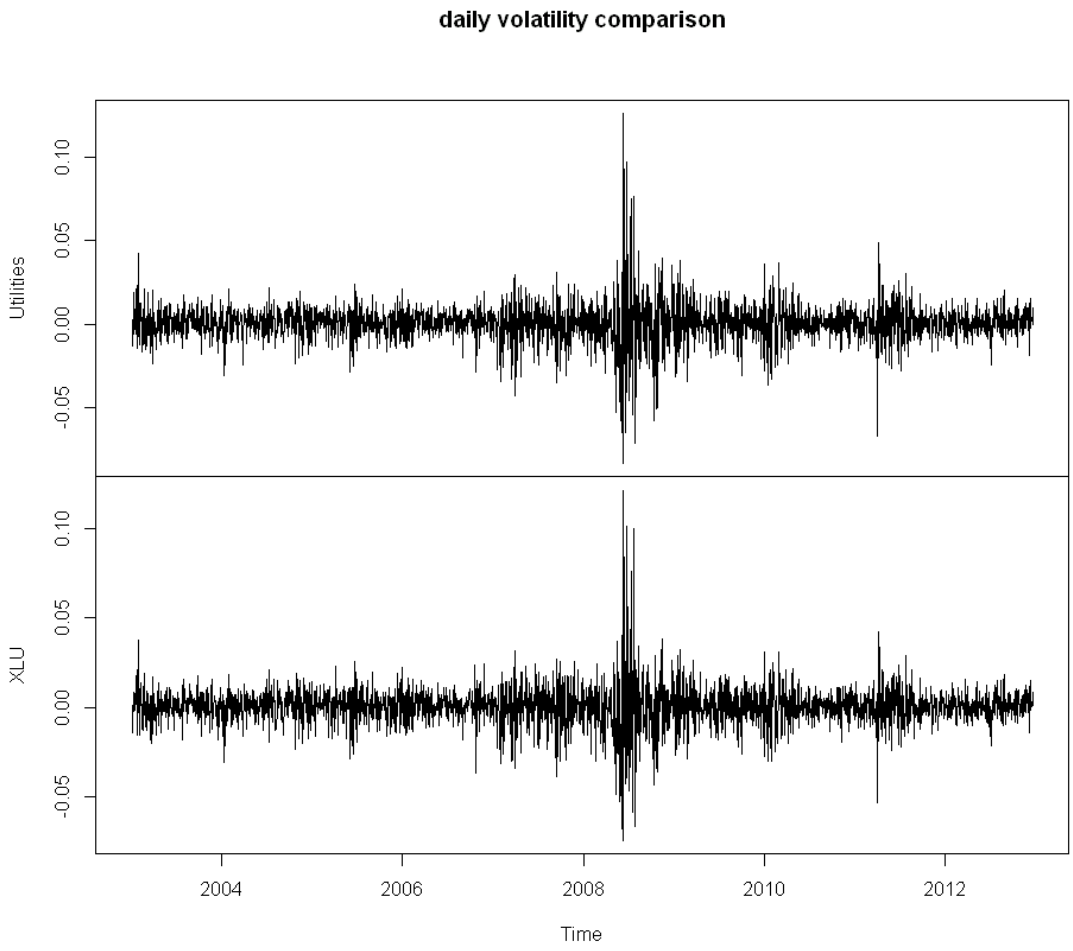

daily volatility comparison
Financials
0.1
0.0
-0.1
-0.2
XLF
0.15
0.10
0.05
-0.05
-0.15
2004
2006
2008
2010
2012
Time

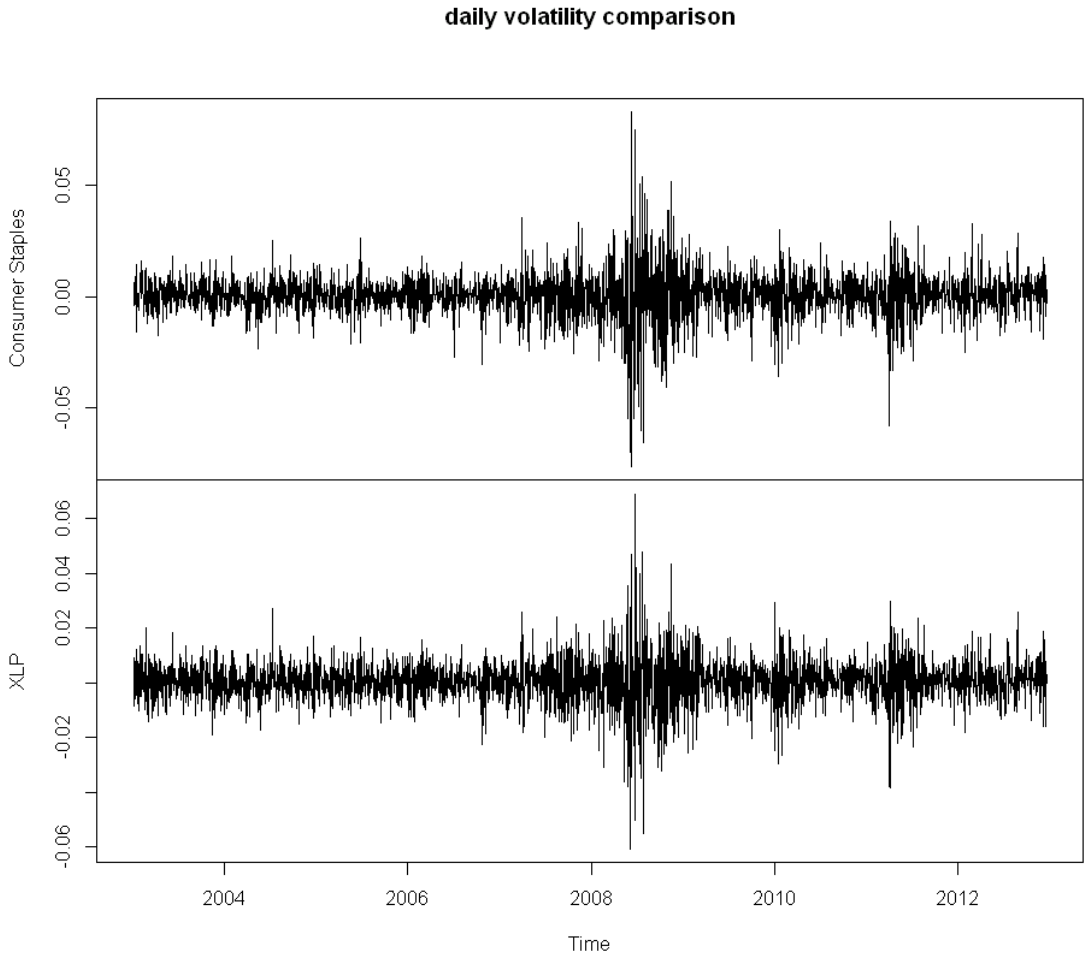

daily volatility comparison
Information Technology
XLK
0.10
0.05
0.00
-0.05
-0.10
2004
2006
2008
2010
2012
Time

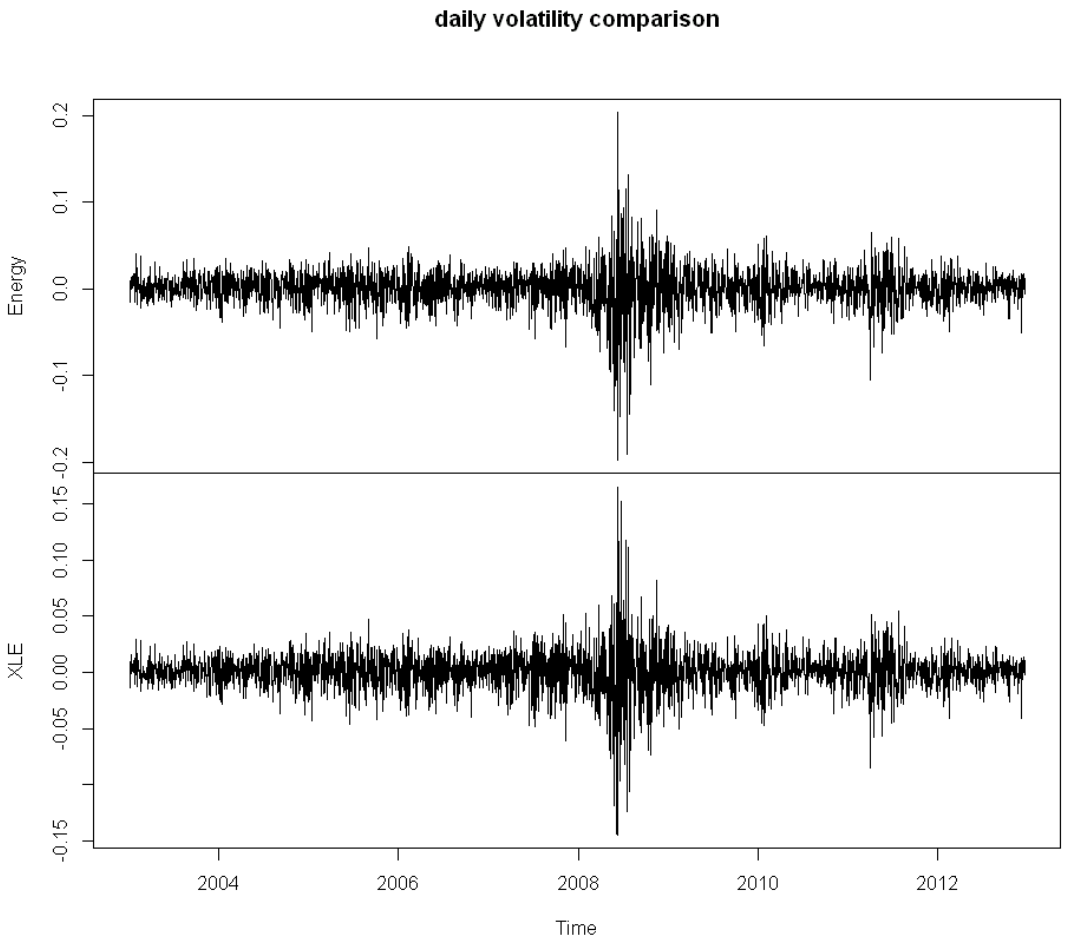

daily volatility comparison
Utilities
XLU
0.10
0.05
0.00
-0.05
2004
2006
2008
2010
2012
Time


daily volatility comparison
Financials
0.1
0.0
-0.1
-0.2
XLF
0.15
0.10
0.05
-0.05
-0.15
2004
2006
2008
2010
2012
Time


daily volatility comparison
Information Technology
XLK
0.10
0.05
0.00
-0.05
-0.10
2004
2006
2008
2010
2012
Time


daily volatility comparison
Utilities
XLU
0.10
0.05
0.00
-0.05
2004
2006
2008
2010
2012
Time


daily volatility comparison
Consumer Staples
0.05
0.00
-0.05
XLP
0.06
0.04
0.02
-0.02
-0.06
2004
2006
2008
2010
2012
Time



We also compared the annualized average daily returns and Sharpe ratios of the 10 sectors with those of the corresponding sector ETFs. Table 1 shows that the performance metrics are comparable albeit some differences.

Table 1: In-sample Performance Comparison of Indices vs. ETFs [1]

Industrials	1.16
Health Care	0.8
Consumer Discretionary	0.41
Financials	-0.82
Information Technology	-0.51
Utilities	0.2
Materials	-0.37
Consumer Staples	0.23
Energy	0.14
Telecommunications Services	-0.24

With these results we are fairly confident that the first principal components are able to give us a good representation of the sectors. Next we proceeded to find the weights of the tangency portfolio. Table 2 lists the weights of the 10 sector indices we constructed. The four negative weights correspond to the four portfolios in the lower right corner of Figure 3, which underperformed the rest of the market in the period we studied and resulted in negative weights in the tangency portfolio.

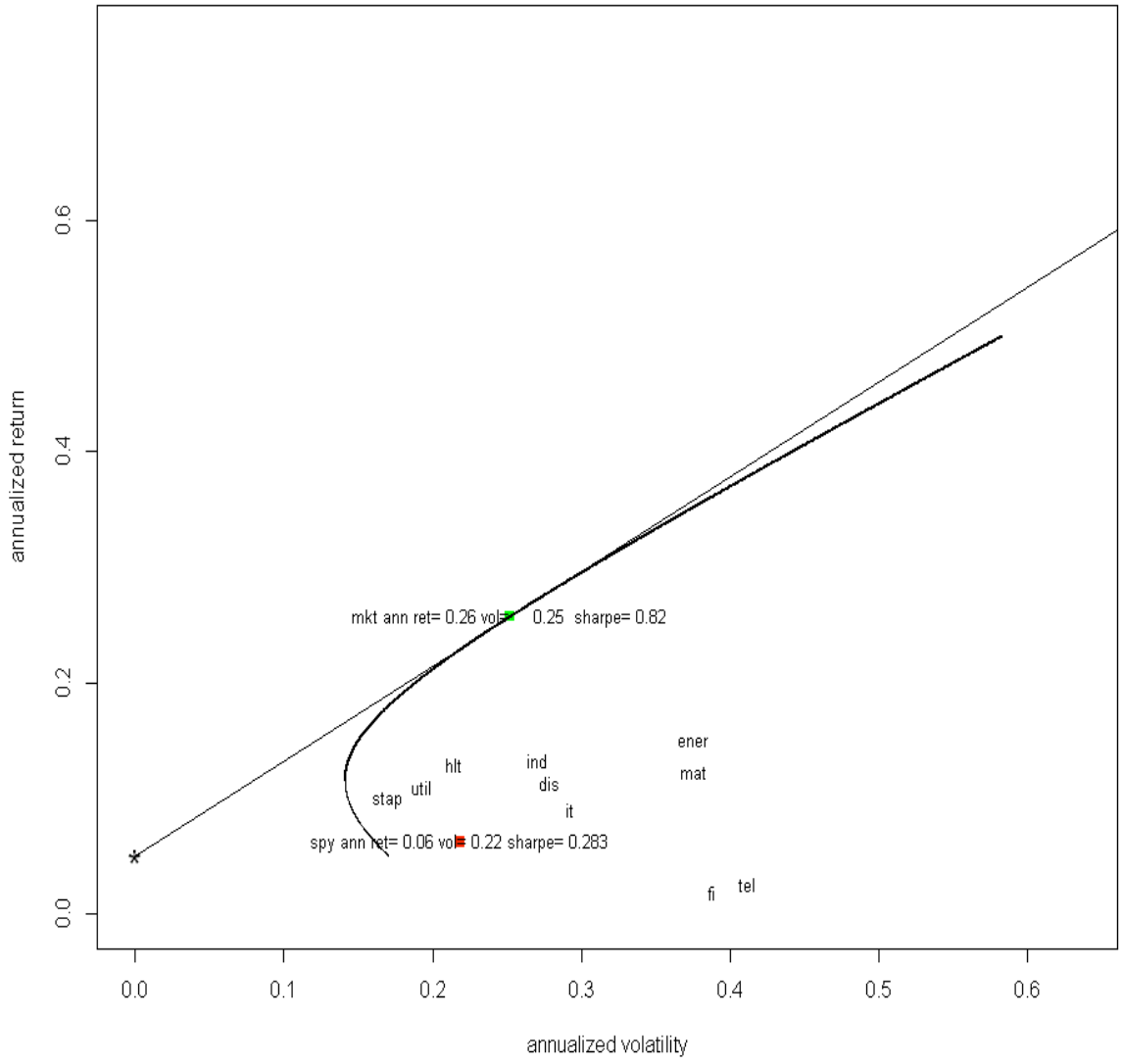
[1] The Sharp ratio difference between the sector index constructed from the first principle component and the sector ETF

Table 2: In-sample tangency portfolio weights

	avg ret yr	Sharpe		avg ret yr	Sharpe
Industrial	0.131	0.419	XLI	0.103	0.456
Health Care	0.119	0.538	XLV	0.066	0.515
Discreptionary	0.103	0.376	XLY	0.102	0.547
Financials	-0.002	-0.006	XLF	0.041	0.186
Technology	0.092	0.279	XLK	0.102	0.606
Utilities	0.104	0.512	XLU	0.117	0.683
Materials	0.157	0.367	XLB	0.124	0.609
Staples	0.088	0.49	XLP	0.098	0.884
Energy	0.154	0.429	XLE	0.182	0.806
Telecomm	-0.059	-0.162	IYZ	0.07	0.373

Figure 3 shows that our constructed market portfolio (indicated by the green square) has a Sharpe ratio of 0.82 and annualized return of 26%, which outperforms SPY with Sharpe ratio 0.28 and return 6%.

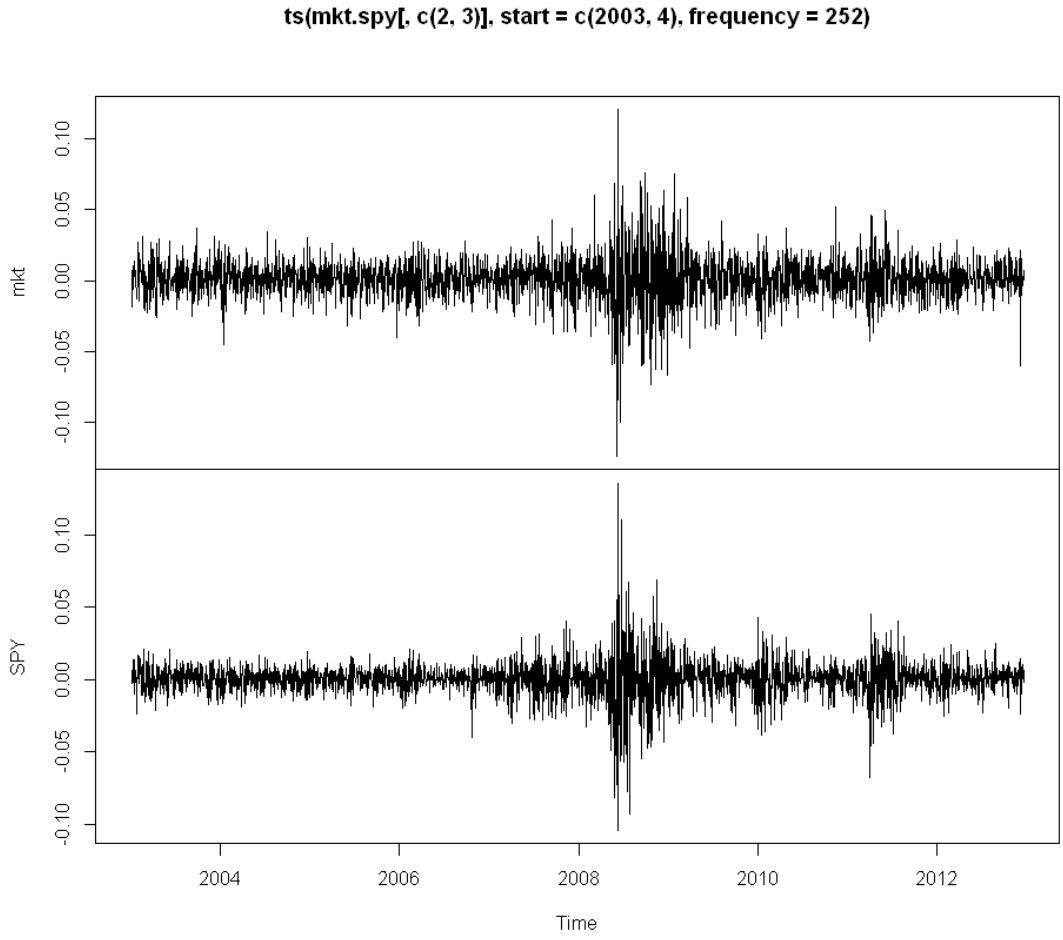
Figure 3: Constructed in-sample tangency portfolio



Additionally, we study the daily returns characteristics of our constructed market portfolio. Figure 4 illustrates that our market portfolio has similar volatility clustering characteristics as that of SPY. The ACF of the return time series in Figure 5 shows that the daily returns are uncorrelated; however the ACF of the returns squared time series indicates correlation. This is known as the "ARCH" effect.

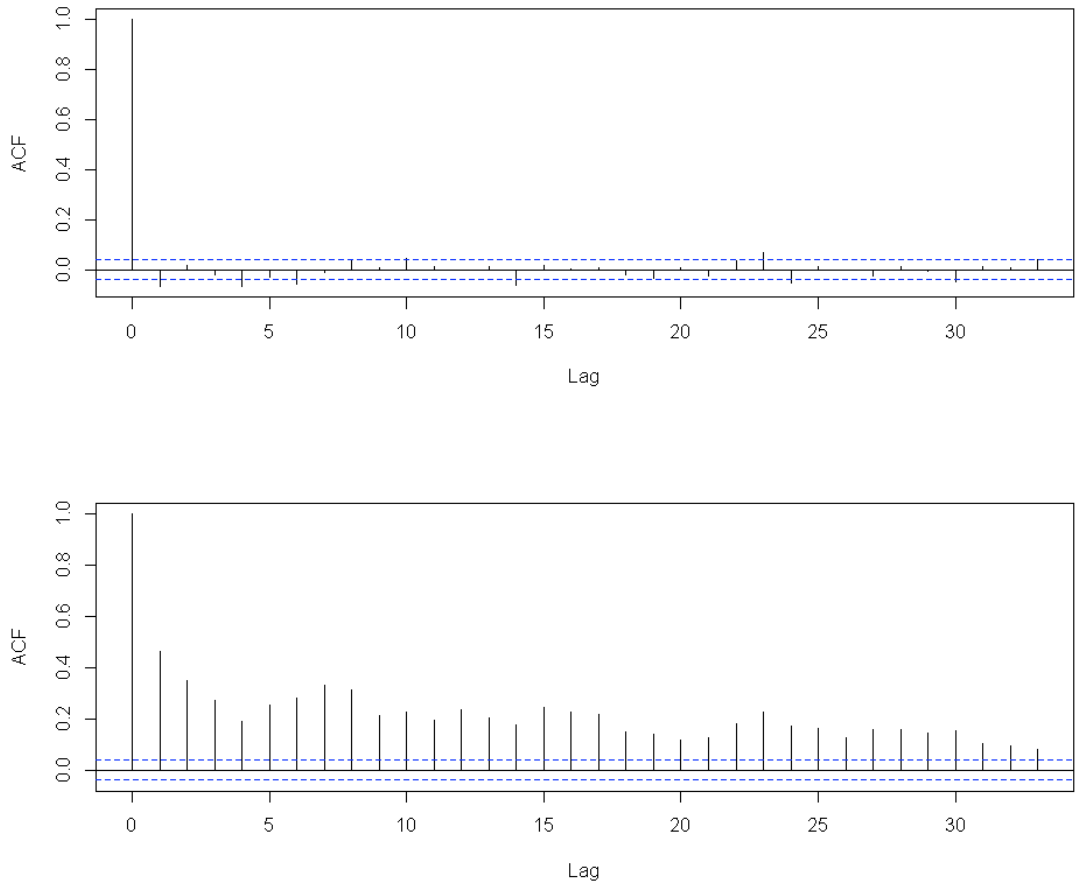
Model 1 in the attachments shows that a GARCH (1,1) volatility model with an ARMA(2,1) return model is a good fit for our return time series. All of the

Figure 4: In-sample tangency portfolio daily returns vs. SPY daily returns



coefficients of the GARCH(1,1) model are significant and the residuals and residuals squared show no correlation.

Figure 5: In-sample Tangency Portfolio Daily Returns ACF



The upper diagram is the return of the time series and the bottom diagram is the ACF of the return series volatility; however the ACF of the returns squared time series indicates correlation.

Out-of-sample analysis

The out-of-sample experiment was conducted using a moving window of training period followed by a window of real trading period using the 10 portfolio weights obtained in the previous period.

In other words, we rebalance the portfolio by performing PCA using the most recent n days of data to construct a new set of 10 sector indices and by solving the tangency portfolio problem using the latest covariance matrix of the new 10 sector indices.

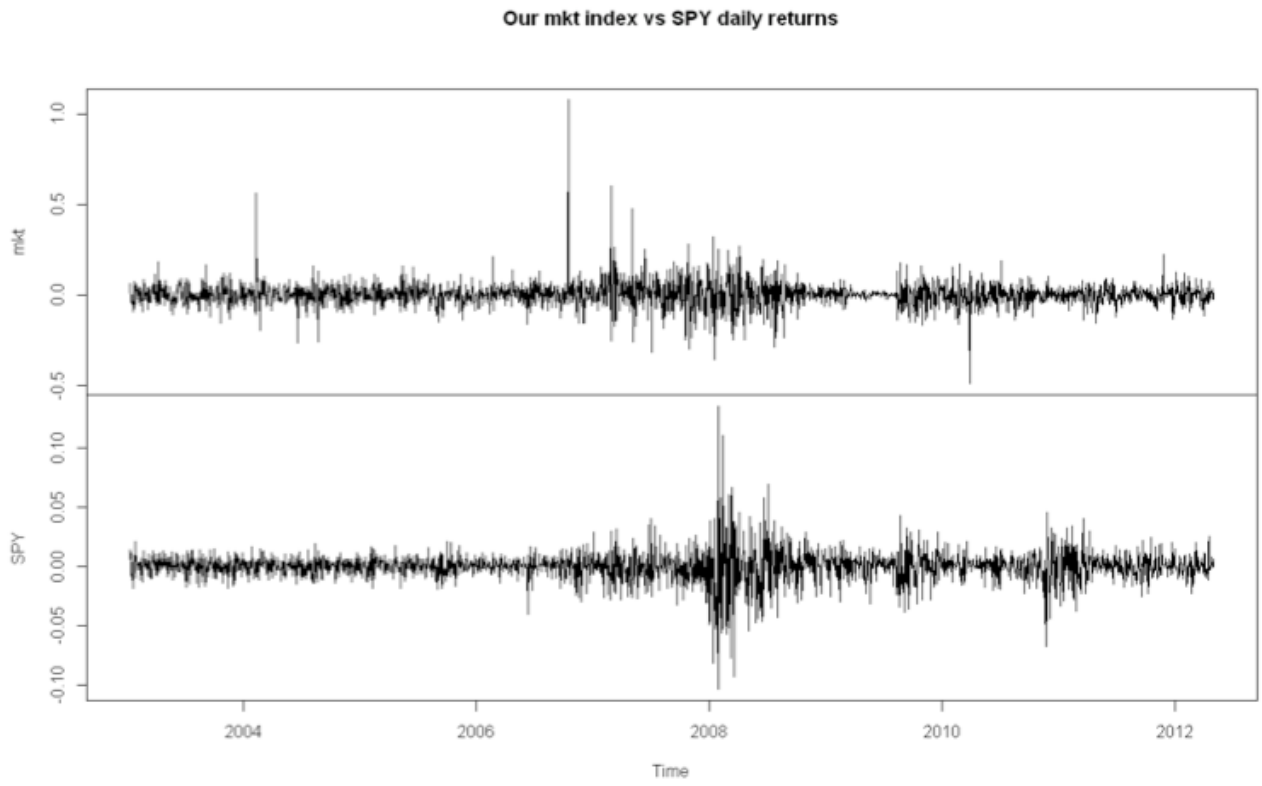
We follow the same procedure for 9 different rebalancing periods ranging from 252 days to 76 days. The summary of the performance results is in Table 3. It appears that the rebalancing period of 98 days yields the best performance.

Our portfolio slightly underperforms SPY in terms of Sharpe ratio (0.21 vs. 0.28). Although the annualized return of our portfolio significantly outperforms SPY (0.24 vs. 0.06), the volatility of our portfolio is much higher than that of SPY (1.14 vs. 0.22).

One possible explanation of this higher volatility is that our portfolio is composed of only the first principal components, which magnify the volatility of the portfolio. The weights of stocks in the first principal component are generally of the same sign, which means these stocks all move in the same direction. Should the 2nd or 3rd principal components be included, our portfolio volatility could decrease because the weights of the 2nd and 3rd components are often of opposite signs and can offset one another. As a result, our Sharpe ratio could also increase.

Figure 6 is a comparison of our market portfolio daily returns time series and that of the SPY. Contrary to the same comparison in the in-sample test, there exist noticeable differences in the volatility patterns. The reason for this could be that the rebalancing somehow decoupled the correlation between our market portfolio and SPY.

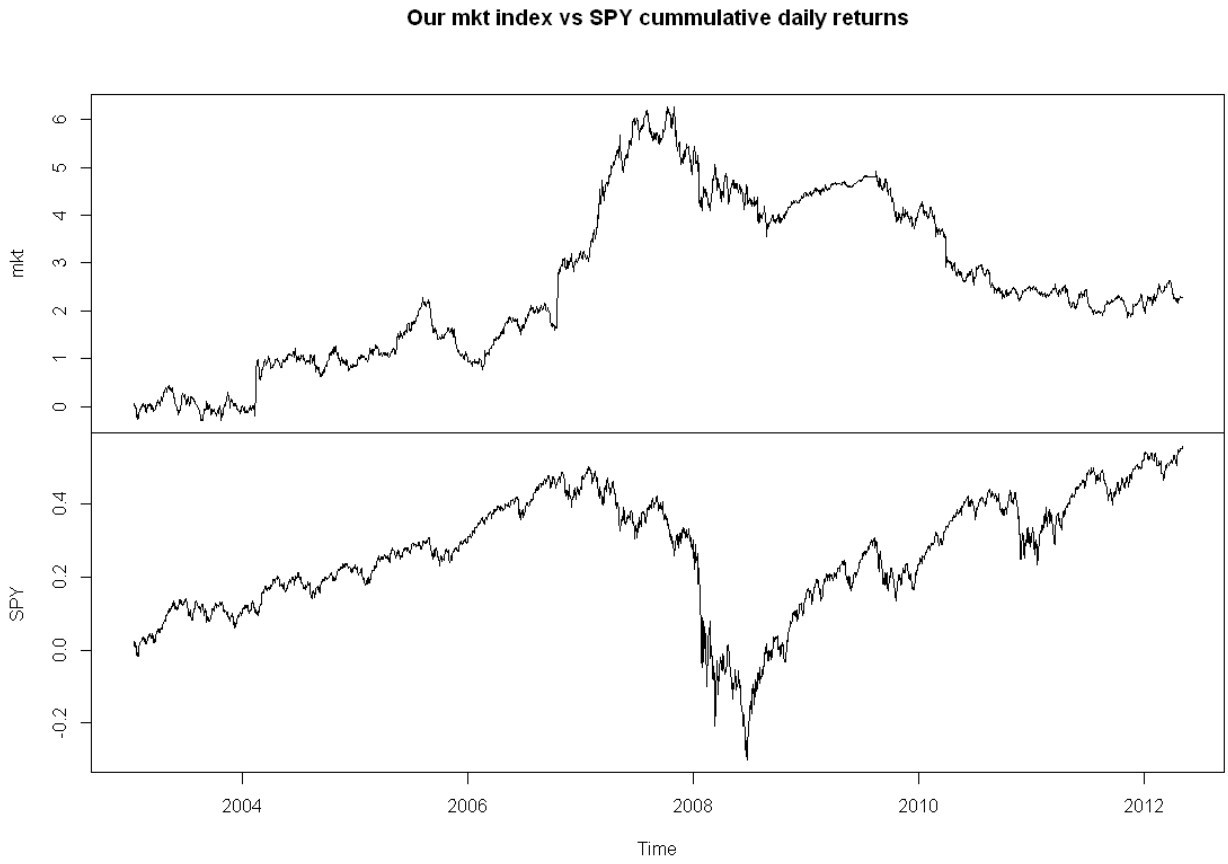
Figure 6: Out-of-sample Tangency Portfolio Daily Returns vs. SPY Daily Returns



The differences between our market portfolio and SPY are also apparent in the cumulative return series in Figure 7. Our market portfolio experienced a large run up from 2004 to 2008, +600% compared to that of SPY, +40%. The SPY suffered a 50%

drop in the 2008 credit crisis.

Figure 7: Out-of-sample Tangency Portfolio Cumulative Daily Returns vs. SPY Cumulative Daily Returns



From Figure 7 we notice that during the financial crisis of 2008, the SPY index dropped dramatically, however compare to SPY our market index only slightly decreased. This is because we took advantage of Markowitz's mean-variance analysis to set up hedging positions (see Table 1) in order to reduce the risk of the entire portfolio.

Figure 8: Out-of-sample tangency portfolio A

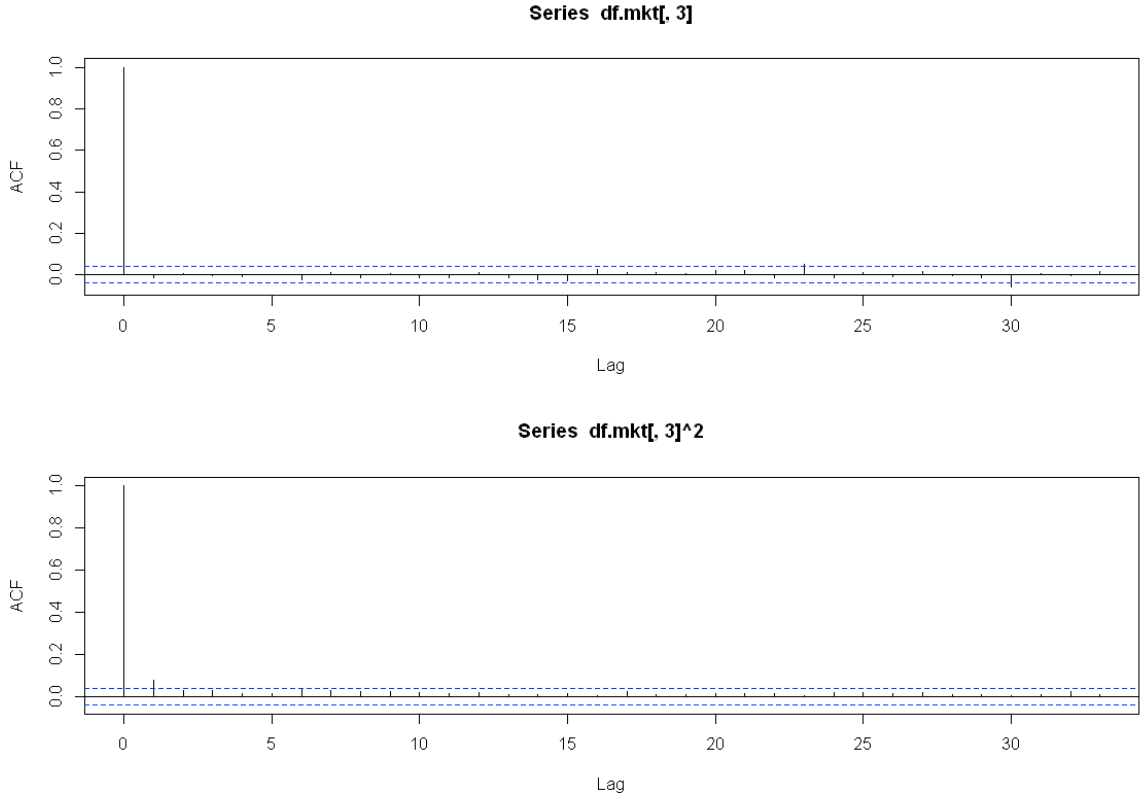


Figure 8 is the ACF of our market portfolio, which shows no correlation in the returns time series, but some correlation in the returns squared time series. Model 2 shows that a GARCH (1,1) with an ARMA(2,1) is a good fit for our market portfolio. These results are in line with what we saw in the in-sample experiment.

Volatility is given by the following ARCH model.

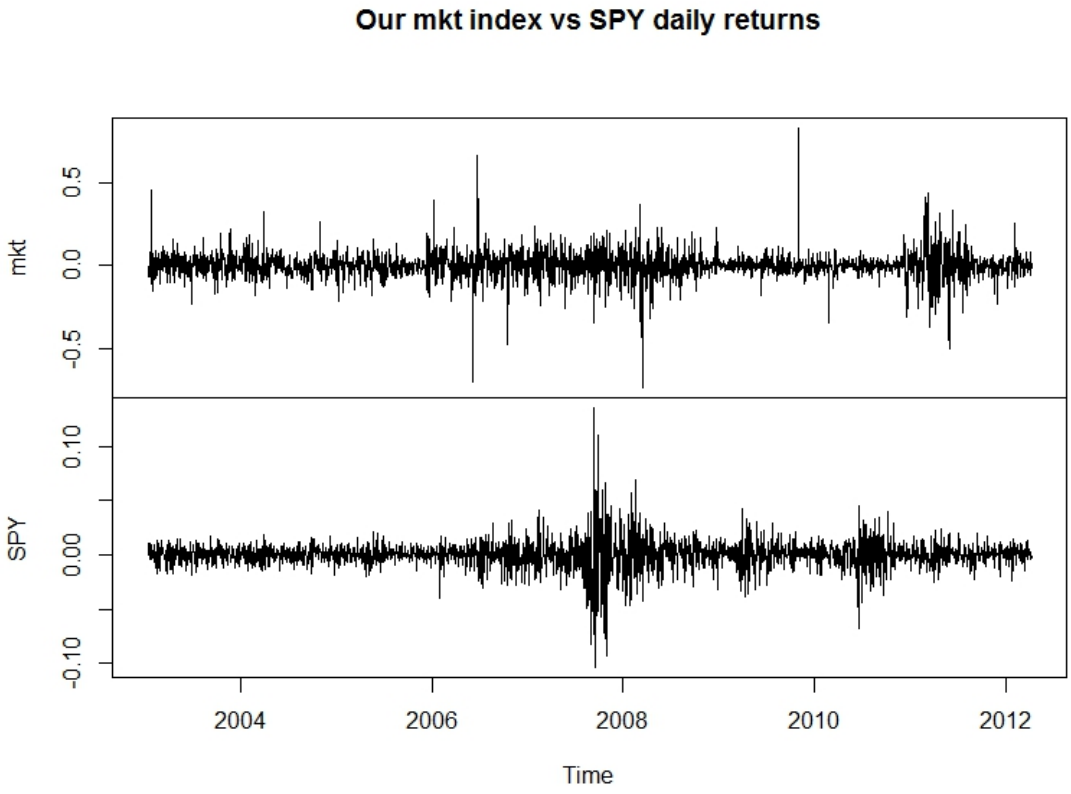
$$r_t = \mu + \sum_{t=1}^{m} \phi_t \, r_{t-i} + \varepsilon_t + \sum_{j=1}^{n} \theta_j \varepsilon_{t-j} + X_t$$

$$X_t = \sigma_t e_t$$

$$\sigma_t{}^2 = a_0 + \sum_{i=1}^{p} \beta_i \sigma_{t-j}^2 + \sum_{j=1}^{q} a_j X_{t-j}^2$$

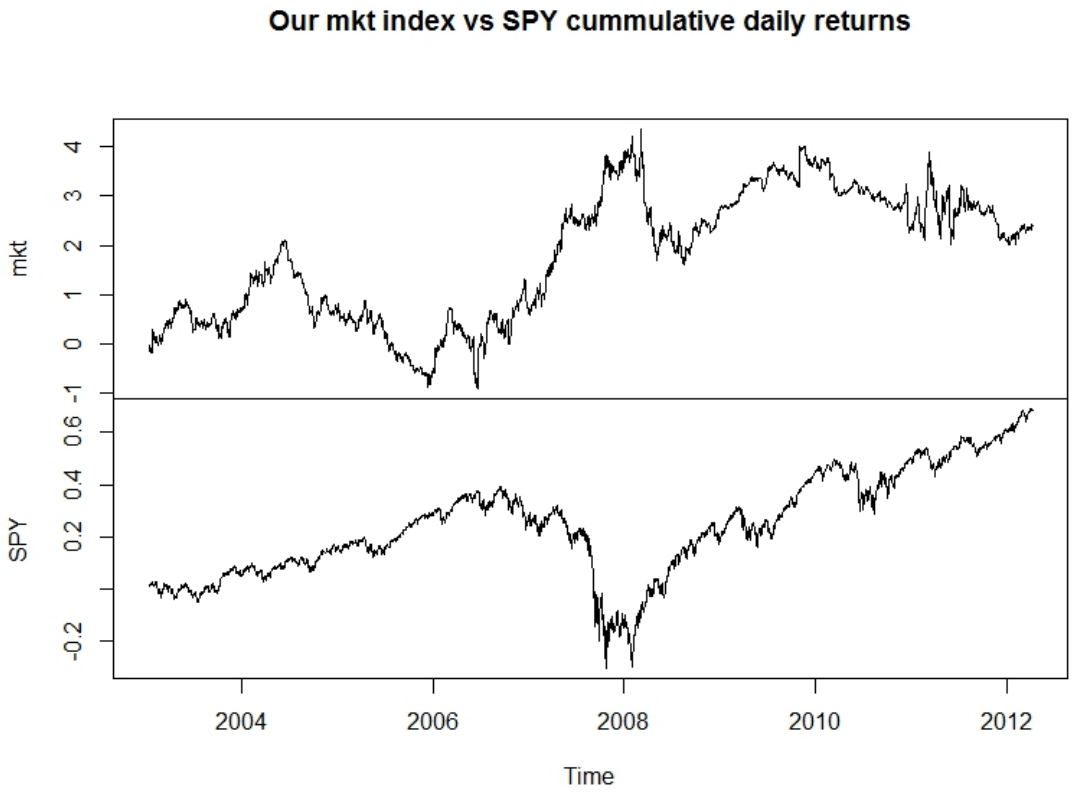
Incorporating the 2nd or 3rd principal components

The out-of-sample results were not as impressive as the in-sample results, but this was to be expected. This could be attributed to the fact that our market portfolio was composed of only the 1st principal components, which magnified the portfolio volatility. This can be mitigated by incorporating the 2nd, or 2nd and 3rd, principal components.



We add the second principal component of each sector to the portfolio and repeat the same tests for 9 different rebalancing periods ranging from 252 days to 76 days. The summary of the performance results is in Table 3. It appears that the

rebalancing period of 186 days yields the best performance. The overall performance is similar to the portfolio constructed from the first principal component. It gained a large run up from 2004 to 2008 but suffered a big drop after 2008. The Sharpe ratio is improved from 0.21 to 0.22.



Conclusion

In this project we took a purely quantitative approach to construct a stock portfolio. This quantitative approach has a theoretical foundation based on principal component analysis. The back testing results showed that this approach could potentially produce market outperforming results.

Our experiments showed that principal component analysis did provide us with indices that were representative of the market. The first principal component also gave us the time series that had the most variance. We were able to construct a tangency portfolio that maximized the Sharpe ratio. The final market portfolio had GARCH (1,1) characteristics with the returns having ARMA (2,1) characteristics. Principle component analysis is very helpful to reduce the dimensionality of a complex financial model, and find the driving factor of the performance of a portfolio.

Future Work

Another possible way to further improve performance is with the help of GARCH volatility prediction. Given a volatility prediction, we can adjust the weights of cash and the market portfolio accordingly to meet the desired return or volatility target.

To illustrate this point, if the predicted GARCH volatility for the coming day is higher than the current volatility, we could decrease our market portfolio weight and increase the cash weight, whereby decreasing the overall portfolio volatility.

Conversely, if the predicted GARCH volatility is lower than the current volatility, we could increase our market portfolio weight and decrease the cash weight, whereby increasing the overall portfolio volatility and increasing the overall return. Since one fund theory advocates that all portfolios on the tangency line are efficient portfolios, by adjusting the weights we are moving up and down the efficient frontier to achieve the desired results.

One key aspect of this strategy is to optimize the parameters of the strategy. Other parts of the system can also be parameterized, such as the number of principal components, the amount of historical data used to do PCA, etc. This creates the need to perform robust back tests on the various combinations of different parameters to find the optimal setting.

References

Santos, Andre A.P. and Moura, Guilherme V. "Dynamic Factor Multivariate GARCH Model" (June 24, 2012). Forthcoming, *Computational Statistics and Data Analysis*. SSRN. Web

"Yahoo! Finance Historical Data." Yahoo! Finance, Web. 2013. <http://finance.yahoo.com/>.

Elton, Edwin J., and Martin J. Gruber. *Modern Portfolio Theory and Investment Analysis*, 4th edition. New York: John Wiley & Sons 1991. Print.

Alexander, C. *Market Models: A Guide to Financial Data Analysis*. Chichester: Wiley, 2001.Print.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. *Times Series Analysis: Forecasting and Control*, Hoboken; Wiley, 2008.Print.

Ruppert, David. *Statistics and Data Analysis for Financial Engineering.* Ithaca: Springer, 2011. Print.

Prigent, Jean-Luc. *Portfolio Optimization and Performance Analysis.* Boca Raton: CRC Press, 2007. Print.

Team Latte, Principle Analysis (PCA) in Quantitative Finance, Risk Latte. Web. February 2012
< http://www.risklatte.com/Articles/QuantitativeFinance/QF_199.php>

Hansen, Peter R., and Asger Lunde. "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?" *Journal of Applied Econometrics* 20.7 (2005): 873-89. JSTOR. Web. 21 Feb. 2012.

Haugh Martin. "Multivariate Distributions and Dimension Reduction Techniques." Web. Spring, 2010.
< http://www.columbia.edu/~mh2078/MultivariateTechniques.pdf>

DeWeese, Jackson Paul. *Markowitz-Style Quartic Optimization for the Improvement of Leveraged ETF Trading,*Electronic Theses and Dissertations. Worcester Polytechnic Institute, 5 Apr. 2013. Web.
< http://www.wpi.edu/Pubs/ETD/Available/etd-042513-204908/unrestricted/jdeweese.pdf>

R Codes

Out-of-sample tangency portfolio daily returns Box test

> Box.test(df.mkt[,3]-mean(df.mkt[,3]),type='Ljung-Box')
Box-Ljung test
data: df.mkt[, 3] - mean(df.mkt[, 3])
X-squared = 0.4804, df = 1, p-value = 0.4882
> Box.test((df.mkt[,3]-mean(df.mkt[,3]))^2,type='Ljung-Box')
Box-Ljung test
data: (df.mkt[, 3] - mean(df.mkt[, 3]))^2 X-squared = 13.681, df = 1, p-value = 0.0002166
Model 2: Out-of-sample tangency portfolio daily returns GARCH(1,1) model
> m=garchFit(df.mkt[,3] ~arma(2,1)+garch(1,1),data=df.mkt[,3],trace=F) > summary(m)
Error Analysis:
Estimate Std. Error t value Pr(>|t|)
mu
ar1
ar2
ma1
omega
alpha1
beta1

Standardised Residuals Tests:

1.066 0.286 -15.591 < 2e-16 *** -4.182 2.89e-05 ***
17.107 < 2e-16 *** 3.996 6.45e-05 ***
8.318 < 2e-16 *** 10.921 < 2e-16 ***
0.0019000 -0.7971537 -0.0992306
0.0017827 0.0511275 0.0237301
0.7552492 0.0004392
0.0441479 0.0001099
0.5498337 0.5508978
0.0661031 0.0504447
Jarque-Bera Test Shapiro-Wilk Test Ljung-Box Test Ljung-Box Test Ljung-Box Test Ljung-Box Test Ljung-Box Test Ljung-Box Test LM Arch Test
Statistic p-Value R Chi^2 191746.2 0
R W 0.8806707 0
R Q(10) 14.27312 0.1608942 R Q(15) 21.93841 0.109429
R
R^2
R^2

R^2 Q(20) 2.219426 0.9999997 R TR^2 0.7481422 0.9999972

```
myWD = "c:/CU/statsMethods"
setwd(myWD)
library(quantmod)
spx = read.table("SP500_companies.csv", header=T, sep=",")
exclude = c('BRK.B')
spx = spx[! spx$Ticker.symbol %in% exclude,]
start = 20030426
minLength = 2520
daysAgo = Sys.Date()-1260
daysAgo = as.numeric(substr(as.character(daysAgo), 1, 4)) * 10000 +
as.numeric(substr(as.character(daysAgo), 6, 7)) * 100 +
as.numeric(substr(as.character(daysAgo), 9, 10))
sectors = unique(spx$GICS.Sector)
all = NULL

# get prices
for (t in spx$Ticker.symbol){
  print(t)
  z <- try(df<-
read.csv(paste("http://ichart.finance.yahoo.com/table.csv?s=",t,"&i
gnore=.csv",sep=""), stringsAsFactors=F))
  if (class(z) == "try-error"){
    print(paste("can't download ",t,sep=""))
    next
  }
  if (nrow(df) < minLength){print("Not enough data.
Skipped.");next}
  df$Date =
as.integer(paste(substr(df$Date,1,4),substr(df$Date,6,7),substr(df$
Date,9,10),sep=""))
  if (nrow(df[df$Date>daysAgo,]) == 0){print("Delisted.
Skipped.");next}
  df = df[,c('Date','Adj.Close')]
  names(df)[2] = t
  if (is.null(all)){
    all = df
  } else {
    all = merge(all,df,by='Date',all.x=T,all.y=T)
  }
}

# get risk free rates
irx =
read.csv(paste("http://ichart.finance.yahoo.com/table.csv?s=^IRX&ig
nore=.csv",sep=""), stringsAsFactors=F)
irx = irx[,c('Date','Adj.Close')]
```

```
irx$Date =
as.integer(paste(substr(irx$Date,1,4),substr(irx$Date,6,7),substr(i
rx$Date,9,10),sep=""))
names(irx)[2] = 'IRX'
all = merge(all, irx, by = 'Date', all.x=T, all.y=T)

# get spy
spy=
read.csv(paste("http://ichart.finance.yahoo.com/table.csv?s=SPY&ign
ore=.csv",sep=""), stringsAsFactors=F)
spy = spy[,c('Date','Adj.Close')]
spy$Date =
as.integer(paste(substr(spy$Date,1,4),substr(spy$Date,6,7),substr(s
py$Date,9,10),sep=""))
names(spy)[2] = 'SPY'
all = merge(all, spy, by = 'Date', all.x=T, all.y=T)
all = all[all$Date >= start,]
all[2:nrow(all),2:ncol(all)] =
all[2:nrow(all),2:ncol(all)]/all[1:(nrow(all)-1),2:ncol(all)]-1
all[1,2:ncol(all)] = NA
all[,2:ncol(all)] = log(all[,2:ncol(all)]+1)
all.whole = all

# start backtesting
rebalance = c(252,230,208,186,164,142,120,98,76)
df.mkt = NULL
for (rebal in rebalance){
  lookback = rebal
  rebal.num = ceiling((nrow(all.whole)-lookback+1)/rebal)
  w.ts = mat.or.vec(10,rebal.num)
  for (k in 1:rebal.num){
    print(paste("rebal=",rebal,"rebal.num=",k))
    all = all.whole[((k-1)*rebal+1):((k-1)*rebal+lookback),]

    ## PCA
    pca = NULL
    par(mfrow=c(5,2))
    for (s in sectors){
      sector = all[,names(all) %in%
spx[spx$GICS.Sector==s,'Ticker.symbol']]
      sector$Date = all[,'Date']
      sector = na.omit(sector)
      pc = princomp(sector[,1:(ncol(sector)-1)])
      wgts = pc$loadings[,1]^2/sum(pc$loadings[,1]^2)
      ts = as.matrix(sector[,1:(ncol(sector)-1)]) %*%
as.matrix(wgts)
      df = data.frame(Date=sector$Date,score=ts)
      names(df)[2] = s
      if (is.null(pca)){
        pca = df
```

```
      } else {
        pca = merge(pca, df, by = 'Date', all.x=T, all.y=T)
      }
    }

    ## Build tangency portolio
    library(quadprog)
    pca = na.omit(pca)
    cov.mat = cov(pca[,2:ncol(pca)])
    sd.vec = sqrt(diag(cov.mat))
    mean.vec = apply(pca[,2:ncol(pca)],2,mean)
    Amat = cbind(rep(1,10),mean.vec)
    muP = seq(0.05,3.50,length=1000)/252
    sdP = muP
    weights = matrix(0,nrow=1000,ncol=10)

    for (i in 1:length(muP)){
      bvec=c(1,muP[i])

result=solve.QP(Dmat=2*cov.mat,dvec=rep(0,10),Amat=Amat,bvec=bvec,m
eq=2)
      sdP[i]=sqrt(result$value)
      weights[i,]=result$solution
    }


irx=read.csv(paste("http://ichart.finance.yahoo.com/table.csv?s=^IR
X&ignore=.csv",sep=""), stringsAsFactors=F)[,c('Date','Adj.Close')]
    irx$Date =
as.integer(paste(substr(irx$Date,1,4),substr(irx$Date,6,7),substr(i
rx$Date,9,10),sep=""))
    mufree = irx[irx$Date==pca[nrow(pca),1],'Adj.Close']/252
    par(mfrow=c(1,1))
    sharpe=(muP-mufree)/sdP
    ind=(sharpe==max(sharpe))
    w=round(weights[ind,],4)
    names(w) = sectors
    print(w)

    ## construct market portfolio
    w.ts[,k] = as.matrix(w)
    if (k == 1) next
    else  mkt = as.matrix(pca[1:rebal,2:11]) %*% w.ts[,k-1]

    if (is.null(df.mkt)) df.mkt =
data.frame(Rebal=rebal,Date=pca$Date[1:rebal],mkt)
    else df.mkt =
rbind(df.mkt,data.frame(Rebal=rebal,Date=pca$Date[1:rebal],mkt))
  }
}
```

```
tmp=by(df.mkt,df.mkt$Rebal,
function(x){x=na.omit(x);return(data.frame(ret=mean(x[,3])*252,sd=s
d(x[,3])*sqrt(252),sharpe=mean(x[,3])/sd(x[,3])*sqrt(252)))})
p=do.call(rbind,as.list(tmp))
r=as.integer(rownames(p[p$sharpe==max(p$sharpe),]))
df.mkt=df.mkt[df.mkt$Rebal==r,]
df.mkt=na.omit(df.mkt)
mkt.spy = merge(df.mkt, all.whole[,c('Date','SPY')], by ='Date')
plot(ts(mkt.spy[,c(3,4)],start=c(2003,9),frequency=252),main='Our
mkt index vs SPY daily returns')
plot(ts(cumsum(mkt.spy[,c(3,4)]),start=c(2003,9),frequency=252),mai
n='Our mkt index vs SPY cummulative daily returns')
mean(mkt.spy$SPY)/sd(mkt.spy$SPY)*sqrt(252)
mean(mkt.spy$mkt)/sd(mkt.spy$mkt)*sqrt(252)

## Fit GARCH
par(mfrow=c(2,1))
acf(df.mkt[,3])
acf(df.mkt[,3]^2)
Box.test(df.mkt[,3]-mean(df.mkt[,3]),type='Ljung-Box')
Box.test((df.mkt[,3]-mean(df.mkt[,3]))^2,type='Ljung-Box')
m=garchFit(df.mkt[,3]
~arma(2,1)+garch(1,1),data=df.mkt[,3],trace=F)
summary(m)
predict(m,n.ahead=3)
```