Assignment #6: Principal Components in Predictive Modeling

Harini K. Anand

1. Introduction

Data for financial assets such as stocks, bonds, commodities, currencies, and cash tend to be highly correlated (that is, they move in relation to each other). For example, the performance of small-cap stocks relative to large-cap stocks is correlated by the overall market sentiment. Principal component analysis (PCA) is a statistical method of dimension reduction that allows us to understand if there are a small number of uncorrelated parts of the data that can explain a large part of the data. It is also used as a remedial measure for multicollinearity in linear regression.

In this report, we present the results of the principal component analysis conducted on the stock portfolio data to predict the value of the Vanguard large-cap index fund. Log return transformations were applied to the raw data for use in the analysis. The first objective was to use PCA for dimension reduction. In this context, PCA was utilized as an unsupervised learning method to identify the principal components to keep. Using an unsupervised decision rule, we kept the first eight principal components. These first eight principal components were used to fit a linear regression model. This is to address the multicollinearity issues in the data. Predictive modeling framework with 70/30 train/test split of the component score data was used to perform the fitting. Upon fitting the model with the first eight principal components, we performed a comparison with two naïve models (a small model and a full model) using the predictive accuracy metric MAE (Mean Absolute Error). The results of the comparison are presented in the report. Lastly, we used the automatic variable selection (backward variable selection method) as a supervised approach for selecting the number of principal components using the component scores data. The resultant model using the backward variable selection method was compared with the linear model fitted the PCA's first eight components and the naïve models using the predictive accuracy metric, MAE.

The report is structured first to describe the data preparation, followed by exploratory data analysis (EDA) using statistical graphs/data visualization tools and the models. Following that, the report contains the results of the principal component analysis and the selection of the number of principal components for use in the linear regression model. The report presents a summary of the fitted linear regression model with the first eight principal components conducted using the predictive modeling framework of train/test split of the component scores data. After that, we showed the model comparison results using the linear model fitted with the first eight principal components and the naïve models (a small model and a full model). Towards the end, the report contains the results of the analysis of the automatic variable selection for selecting the principal components and the comparison results of the backward variable selection model with the previous models.

2. Data

The dataset used for this analysis is the stock portfolio data. It contains the time series data of daily closing stock prices for twenty stocks and a large-cap index fund from Vanguard from 3-Jan-2012 to 31-Dec-2013 for 21 stocks (502 observations).

3. Data preparation

As part of the data preparation, the date attribute in the raw dataset was transformed to the R Date format (YYYY-mm-dd). The dataset was then sorted using the transformed date. We then generated log-returns of the individual stocks. Log-return for a stock is obtained by taking the natural logarithm of the future value to the present value. Since stock prices behave similar to exponential functions, to make them close to a normally distributed variable, log returns are used which have more stable distributions than the arithmetic returns.

Code Snippet 01 in section 12.1 contains the R code for the data preparation.

4. Exploratory data analysis using statistical graphs and data visualization tools

In this section, we present the results of the exploratory data analysis conducted by examining the correlations among all log-return values of the stocks. For that, we computed the full correlation matrix using all the log-return stock variables including the VV index fund. From the full correlation matrix, we extracted the last column which contained the correlation coefficients of the log-return stock variables computed against the log-return values of the VV index fund (shown in Table 01). The same is also plotted using as a bar graph in FIG 01.

Log-return variable	Correlation Coefficient
AA	0.6324106
BAC	0.6501877
BHI	0.5774988
CVX	0.7209041
DD	0.6895190
DOW	0.6264550
DPS	0.4435005
GS	0.7121620
HAL	0.5974989
HES	0.6107960
HON	0.7683784
HUN	0.5819449
JPM	0.6578478
KO	0.5997988
MMM	0.7608489
MPC	0.4731198
PEP	0.5075264
SLB	0.6928534
WFC	0.7335731
XOM	0.7211079
VV	1.0000000

Table 01: Correlation Coefficient for each of the log-return stocks against the log-return VV index fund.

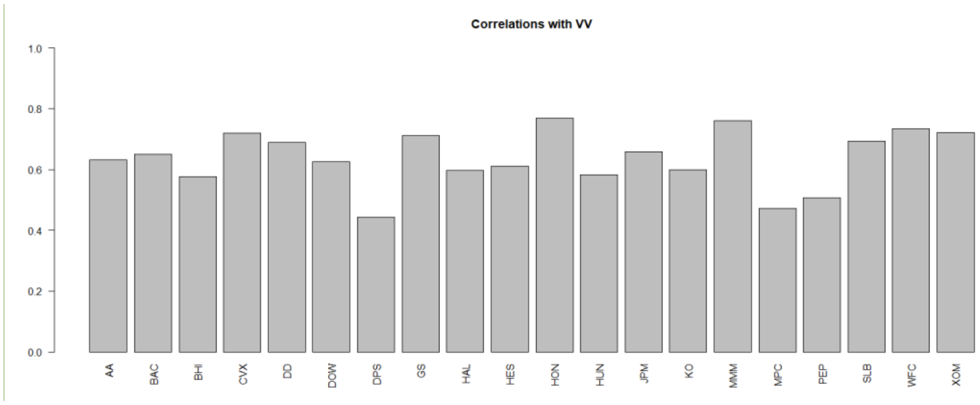


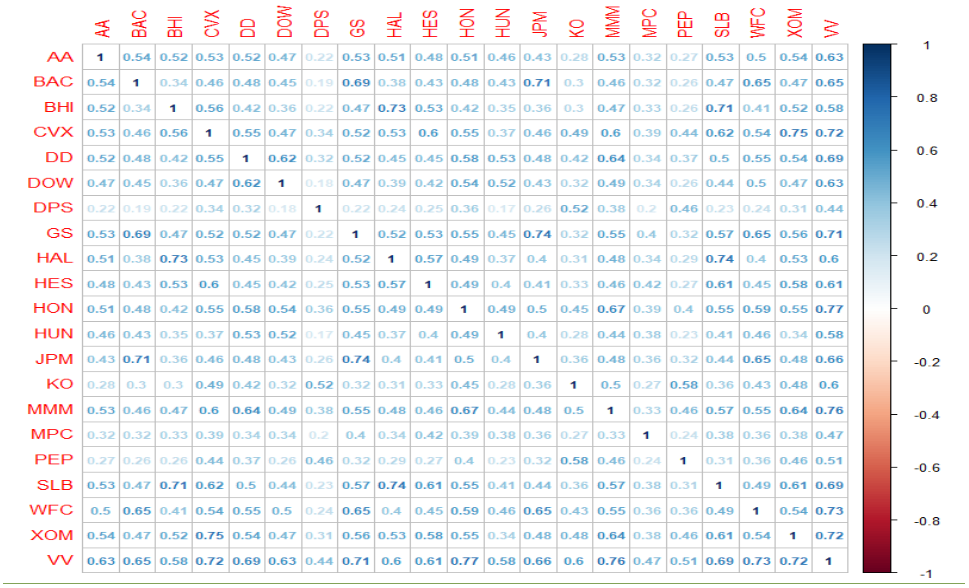
FIG 01: A plot of the correlation coefficients for each of the log-return stocks against the log-return VV index fund.

From Table 01 and FIG 01, we determined that HON and MMM have the top two high correlation coefficients with respect to VV, with the values 0.7683784 and 0.7608489 respectively.

To be able to view all the pairwise correlations in the data, we created a corrplot. From the corrplot, we can not only determine the pairwise correlations with the VV index fund, but also the correlation coefficients among the other stock variables. That is the advantage of the corrplot (FIG 02) over the simple bar chart (FIG 01). We can detect the multicollinearity by examining the corrplot for correlation coefficients that are closers to -1 or +1. In the stock portfolio data, all the correlation coefficients are positive values.

DPS, MPC, PEP are three variables that have low VIF values (most of the correlation coefficient are below or about 0.50).

VV, SLB, WFC, XOM are some variables that have high VIF values (most of the correlation coefficients are above 0.50)



	AA	BAC	BHI	CVX	DD	DOW	DPS	GS	HAL	HES	HON	HUN	JPM	KO	MMM	MPC	PEP	SLB	WFC	XOM	VV
AA	1	0.54	0.52	0.53	0.52	0.47	0.22	0.53	0.51	0.48	0.51	0.46	0.43	0.28	0.53	0.32	0.27	0.53	0.5	0.54	0.63
BAC	0.54	1	0.34	0.46	0.48	0.45	0.19	0.69	0.38	0.43	0.48	0.43	0.71	0.3	0.46	0.32	0.26	0.47	0.65	0.47	0.65
BHI	0.52	0.34	1	0.56	0.42	0.36	0.22	0.47	0.73	0.53	0.42	0.35	0.36	0.3	0.47	0.33	0.26	0.71	0.41	0.52	0.58
CVX	0.53	0.46	0.56	1	0.55	0.47	0.34	0.52	0.53	0.6	0.55	0.37	0.46	0.49	0.6	0.39	0.44	0.62	0.54	0.75	0.72
DD	0.52	0.48	0.42	0.55	1	0.62	0.32	0.52	0.45	0.45	0.58	0.53	0.48	0.42	0.64	0.34	0.37	0.5	0.55	0.54	0.69
DOW	0.47	0.45	0.36	0.47	0.62	1	0.18	0.47	0.39	0.42	0.54	0.52	0.43	0.32	0.49	0.34	0.26	0.44	0.5	0.47	0.63
DPS	0.22	0.19	0.22	0.34	0.32	0.18	1	0.22	0.24	0.25	0.36	0.17	0.26	0.52	0.38	0.2	0.46	0.23	0.24	0.31	0.44
GS	0.53	0.69	0.47	0.52	0.52	0.47	0.22	1	0.52	0.53	0.55	0.45	0.74	0.32	0.55	0.4	0.32	0.57	0.65	0.56	0.71
HAL	0.51	0.38	0.73	0.53	0.45	0.39	0.24	0.52	1	0.57	0.49	0.37	0.4	0.31	0.48	0.34	0.29	0.74	0.4	0.53	0.6
HES	0.48	0.43	0.53	0.6	0.45	0.42	0.25	0.53	0.57	1	0.49	0.4	0.41	0.33	0.46	0.42	0.27	0.61	0.45	0.58	0.61
HON	0.51	0.48	0.42	0.55	0.58	0.54	0.36	0.55	0.49	0.49	1	0.49	0.5	0.45	0.67	0.39	0.4	0.55	0.59	0.55	0.77
HUN	0.46	0.43	0.35	0.37	0.53	0.52	0.17	0.45	0.37	0.4	0.49	1	0.4	0.28	0.44	0.38	0.23	0.41	0.46	0.34	0.58
JPM	0.43	0.71	0.36	0.46	0.48	0.43	0.26	0.74	0.4	0.41	0.5	0.4	1	0.36	0.48	0.36	0.32	0.44	0.65	0.48	0.66
KO	0.28	0.3	0.3	0.49	0.42	0.32	0.52	0.32	0.31	0.33	0.45	0.28	0.36	1	0.5	0.27	0.58	0.36	0.43	0.48	0.6
MMM	0.53	0.46	0.47	0.6	0.64	0.49	0.38	0.55	0.48	0.46	0.67	0.44	0.48	0.5	1	0.33	0.46	0.57	0.55	0.64	0.76
MPC	0.32	0.32	0.33	0.39	0.34	0.34	0.2	0.4	0.34	0.42	0.39	0.38	0.36	0.27	0.33	1	0.24	0.38	0.36	0.38	0.47
PEP	0.27	0.26	0.26	0.44	0.37	0.26	0.46	0.32	0.29	0.27	0.4	0.23	0.32	0.58	0.46	0.24	1	0.31	0.36	0.46	0.51
SLB	0.53	0.47	0.71	0.62	0.5	0.44	0.23	0.57	0.74	0.61	0.55	0.41	0.44	0.36	0.57	0.38	0.31	1	0.49	0.61	0.69
WFC	0.5	0.65	0.41	0.54	0.55	0.5	0.24	0.65	0.4	0.45	0.59	0.46	0.65	0.43	0.55	0.36	0.36	0.49	1	0.54	0.73
XOM	0.54	0.47	0.52	0.75	0.54	0.47	0.31	0.56	0.53	0.58	0.55	0.34	0.48	0.48	0.64	0.38	0.46	0.61	0.54	1	0.72
VV	0.63	0.65	0.58	0.72	0.69	0.63	0.44	0.71	0.6	0.61	0.77	0.58	0.66	0.6	0.76	0.47	0.51	0.69	0.73	0.72	1

FIG 02: A plot of the correlation coefficients among all the log-return stock variables (including the log-return VV index fund).

A statistical graph is a pictorial representation to describe and quantify the characteristics of data (e.g., bar charts, pie chart, histogram etc.) for statistical evaluation. On the other hand, data visualization is a broader term that is used to present data in a visual context for the exploration of the patterns, trends, outliers, and correlations in the data set.

Code Snippet 02 in section 12.2 contains the R code for the Exploratory Data Analysis (EDA) using statistical graphs and data visualization plots.

5. Exploratory data analysis using models

In this section, we present two naïve models – a small model and a full model. The purpose of the two models is to compute the VIF values of all the variables for the examination of the multicollinearity among the variables.

The small model is obtained by fitting the log return stock values of the variables GS, DD, DOW, HON HUN, JPM, KO, MMM, XOM to predict the log return value of the VV index fund.

The full model is obtained by fitting all the log return stock values (namely AA, BAC, GS, JPM, WFC, BHI, CVX, DD, DOW, DPS, HAL, HES, HON, HUN, KO, MMM, MPC, PEP, SLB, and XOM) to predict the log return value of the VV index fund.

The fitting for the small model results in the following equation:

VV = 0.0001008 + 0.0784765 * GS + 0.0354057 * DD + 0.0406763 * DOW + 0.1449817 * HON + 0.0385118 * HUN + 0.0505123 * JPM + 0.1419686 * KO + 0.1336002 * MMM + 0.1480728 * XOM

Based on the summary table in Table 02, we can conclude at the level of significance of 0.001, we reject that the joint null hypothesis that all the coefficients have no effect on the overall regression. This is based on the p-value (with a corresponding F-statistic value of 313.5) of less than 2.2e-16. Also, it can be gleaned from Table 02 is that all the individual regression coefficient but the intercept and the coefficient for log return variable DD are strongly significant at the level of significance of 0.001 meaning we can the null hypothesis for these individual regression coefficients to be zero at the level of significance of 0.001. The regression coefficient for the log return variable DD is only significant at a level of significance of 0.05. The intercept is NOT significant. The coefficient of determination (Adjusted R-squared) for the small model is 84.9% indicating that 84.9 percent of the variation in the log-return of VV index fund can be explained by variations in the predictor variables.

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0001008	0.0001331	0.757	0.44929
GS	0.0784765	0.0138277	5.675	2.37e-08 ***
DD	0.0354057	0.0177154	1.999	0.046204 *
DOW	0.0406763	0.0116993	3.477	0.000552 ***
HON	0.1449817	0.0170837	8.487	2.53e-16 ***
HUN	0.0385118	0.0077371	4.978	8.93e-07 ***
JPM	0.0505123	0.0132262	3.819	0.000151 ***
KO	0.1419686	0.0176282	8.054	6.14e-15 ***
MMM	0.1336002	0.0239378	5.581	3.96e-08 ***
XOM	0.1480728	0.0213601	6.932	1.31e-11 ***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002951 on 491 degrees of freedom
Multiple R-squared:  0.8518,         Adjusted R-squared:  0.849
F-statistic: 313.5 on 9 and 491 DF,  p-value: < 2.2e-16

Residuals:
       Min        1Q     Median        3Q       Max
-0.0139179 -0.0016005 -0.0000926  0.0016690  0.0172703
```

Table 02: Regression summary table for the small model - lm(formula = VV ~ GS + DD + DOW + HON + HUN + JPM + KO + MMM + XOM, data = returns.df)

Next, we computed the VIF values for the predictor variables used in the small model. The relationship between the predictor variables can be judged by examining a quantity called the VIF (Variance Inflation Factor). We used a cutoff value of 3 to determine if there is any multicollinearity among the variables. VIF is related to the square of the multiple correlation coefficient (the coefficient of determination, R-squared value) that results when the predictor variable is regressed against all the other predictor variables**. A VIF value of 3 implies the coefficient of determination for the predictor variable when regressed against all the other predictor variables is 66.7% meaning 66.7 percent of the**

variation in the predictor variable is explained by the other predictor variables (an indication of the presence of multicollinearity).

The VIF values for the predictor variables used in the small model in decreasing order are shown in Table 03.

Predictor Variable	VIF value
GS	2.705795
MMM	2.590177
DD	2.368257
JPM	2.324600
HON	2.261397
XOM	2.073721
DOW	1.919773
HUN	1.633336
KO	1.473202

Table 03: VIF values for the predictor variables in the small model.

Since none of the VIF values are greater than 3, there is no indication of multicollinearity among the predictor variables used in the small model.

The fitting for the full model results in the following equation:

VV = 9.953e-05 + 1.538e-02 * AA + 2.723e-02 * BAC + 3.434e-02 * GS + 2.224e-02 * JPM + 7.738e-02 * WFC + 1.604e-02 * BHI + 5.742e-02 * CVX + 1.003e-02 * DD + 3.600e-02 * DOW + 5.659e-02 * DPS -1.976e-03 * HAL + 4.393e-03 * HES + 1.071e-01 * HON + 2.867e-02 * HUN + 9.425e-02 * KO + 1.093e-01 * MMM + 1.079e-02 * MPC + 2.092e-02 * PEP + 4.851e-02 * SLB + 5.797e-02 * XOM

Table 04 shows the regression summary table for the fitted full model.

From Table 04, we can conclude that at a level of significance of 0.001, we can reject the joint null hypothesis that all the coefficients have no effect on the overall regression. This is based on the p-value (F-statistic is 179) of less than 2.2e-16. For the individual coefficient tests, the variables for the log returns of the stocks WFC, DOW, DPS, HON, HUN, KO, MMM, and SLB are strongly statistically significant at the level of significance of 0.001. So, at a level of significance of 0.001, we can reject the null hypothesis based on the test using Student's T-distribution that the individual regression coefficient for the above variables is zero.

The coefficients for BAC and CVX are significant at only at a level of significance of 0.01. The coefficients for GS and XOM are significant at a level of significance of 0.05. JPM is significant at a level of significance 0.1. The rest of the coefficients for the variables AA, BHI, DD, HAL, HES, MPC, PEP, and the intercept are NOT significant.

Also, from the summary table, Table 04, we can determine that the coefficient of determination for the full model is 87.68% indicating 87.68 percent of the variation in the log return of VV (Vanguard index fund) can be explained by the variations in all the predictor variables used in the full model.

Term	Estimate	Std. Error t	Value	Pr(>|t|)
(Intercept)	9.95E-05	1.21E-04	0.82	0.412433
AA	1.54E-02	1.04E-02	1.479	0.139894
BAC	2.72E-02	9.70E-03	2.808	0.005188 **
GS	3.43E-02	1.36E-02	2.529	0.011766 *
JPM	2.22E-02	1.33E-02	1.674	0.094849 .
WFC	7.74E-02	1.58E-02	4.897	1.33e-06 ***
BHI	1.60E-02	1.16E-02	1.382	0.167752
CVX	5.74E-02	2.07E-02	2.776	0.005720 **
DD	1.00E-02	1.63E-02	0.618	0.537144
DOW	3.60E-02	1.07E-02	3.366	0.000824 ***
DPS	5.66E-02	1.49E-02	3.79	0.000170 ***
HAL	-1.98E-03	1.21E-02	-0.163	0.870344
HES	4.39E-03	9.69E-03	0.453	0.650432
HON	1.07E-01	1.61E-02	6.658	7.62e-11 ***
HUN	2.87E-02	7.22E-03	3.969	8.31e-05 ***
KO	9.43E-02	1.85E-02	5.103	4.82e-07 ***
MMM	1.09E-01	2.20E-02	4.964	9.63e-07 ***
MPC	1.08E-02	7.02E-03	1.536	0.125134
PEP	2.09E-02	2.03E-02	1.028	0.304293
SLB	4.85E-02	1.45E-02	3.339	0.000905 ***
XOM	5.80E-02	2.30E-02	2.519	0.012094 *

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002666 on 480 degrees of freedom
Multiple R-squared:  0.8818,  Adjusted R-squared:  0.8768
F-statistic:   179 on 20 and 480 DF,  p-value: < 2.2e-16
Residuals:
       Min        1Q    Median        3Q       Max
-0.0139266 -0.0015542  0.0000313  0.0015042  0.0140759
```

Table 04: Regression summary table for the full model lm(formula = VV ~ AA + BAC + GS + JPM + WFC + BHI + CVX + DD + DOW + DPS + HAL + HES + HON + HUN + KO + MMM + MPC + PEP + SLB + XOM, data = returns.df)

After fitting the model, we computed the VIF (Variance Inflation Factor) values for all the predictor variables in the full model. The objective is to identify the multicollinearity among these variables. If any of the VIF values are greater 3 (as explained in the case of the small model), it indicates the presence of multicollinearity among the variables. Table 05 shows the VIF values sorted in the decreasing order for all the predictor variables in the full model.

In Table 05, the variables SLB and GS have VIF values greater than 3. Therefore, there are concerns about multicollinearity in the data.

We next used the statistical method of principal component analysis to remediate the multicollinearity identified in the data.

Code Snippet 03 in section 12.3 contains the R code for the Exploratory Data Analysis (EDA) using models.

Predictor Variable	VIF value
SLB	**3.258982**
GS	**3.197811**
XOM	2.949826
CVX	2.920187
HAL	2.919924
JPM	2.875959
BAC	2.686535
MMM	2.684942
BHI	2.651368
WFC	2.532626
HON	2.455876
DD	2.441486
HES	2.09606
AA	2.02627
KO	1.981647
DOW	1.965886
HUN	1.743913
PEP	1.720663
DPS	1.525629
MPC	1.376706

Table 05: VIF values for all the predictor variables

6. Variable transformation using the principal component analysis

In this section, we present the results of applying the principal component analysis on the log return values of the predictor variables representing the stock prices. The principal components are NOT standardized (i.e., scaling and mean centering) to mean zero and unit variance. However, in our case, the scale invariance is achieved by utilizing the log-return transformation which has already been applied to the stock prices. Due to that, the principal component calculation is done using the ***princomp()*** R function with the ***cor*** logical value set to TRUE (meaning by using the correlation matrix).

The loading in the principal component analysis provides the correlation between a principal component and the variable. It is weight by which each variable should be multiplied to get the principal component score.

Table 06 shows the loadings of the first and the second principal components obtained by using the **princomp()** function on the log-return data.

Predictor variable	PC1 loading value	PC2 loading value
AA	0.2289477	0.15741300
BAC	0.2246401	0.20081224
BHI	0.2186159	0.15017188
CVX	0.2532034	-0.09115728
DD	0.2414028	-0.03433672
DOW	0.2143027	0.09442208
DPS	0.1386150	-0.52040357
GS	0.2507545	0.19111046
HAL	0.2281190	0.14600696
HES	0.2284285	0.10831601
HON	0.2479589	-0.06892393
HUN	0.1974712	0.14011003
JPM	0.2273526	0.10656303
KO	0.1881999	-0.47871023
MMM	0.2514868	-0.13960716
MPC	0.1711442	0.04823903
PEP	0.1702056	-0.48759946
SLB	0.2497330	0.13757414
WFC	0.2428197	0.06010682
XOM	0.2535416	-0.08084532

Table 06: shows the loading values of pc.1 and pc.2 for each of the predictor variable

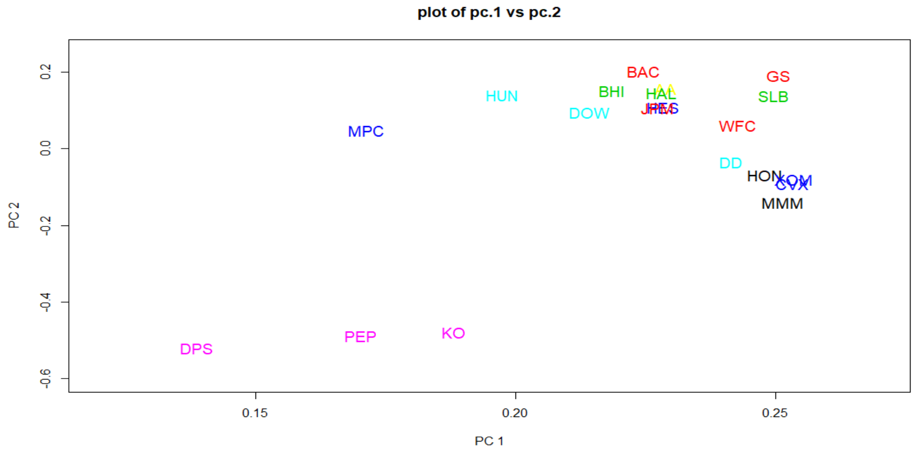


FIG 03: Plot of pc.1 and pc.2 loadings for each of the predictor variable with the labeled stock ticker. (The color of the font for the predictor variable is based on the industry the stock belongs to)

In FIG 03, we notice a few clusters/groupings. One cluster consists of the stock tickers AA and HAL. Close to it are also the stock tickers HES and JPM superimposed. A second cluster contains XOM and CVX (In this case, both are oil refining stocks).

Code Snippet 04 in section 12.4 contains the R code for the principal component analysis.

7. Principal component selection

In this section, we present the decision rules that were evaluated for keeping the principal components. We selected the scree plot decision rule and applied it to select the principal components from the twenty principal components created by the **princomp()** R function.

The decision rules for the principal component selection that were evaluated are:

1. Just enough principal components are kept to explain some percentage of the total variation of the variables.
2. Only those principal components are kept whose eigenvalues are less than the average eigenvalue. The average eigenvalue is the average variance of all the original variables. This rule allows to keep those components that account for more variance than the average for the observed variables.
3. When the components are extracted from the correlation matrix (the data is standardized (transformed) to mean zero and unit variance), the average variance is one. Principal components with eigenvalues less than one are excluded. Typically, a cutoff or threshold (e.g. 0.7) is used.
4. Keep the principal components based on the examination of the plot of eigenvalue of the principal component against the index of the principal component called scree plot. The scree plot shows the faction of the total variance in the data as represented by each principal component.
5. A variation of the scree diagram is the log-eigenvalue diagram consisting of a plot of logarithm value of the eigenvalue of the principal component for each principal component.

For our analysis, among the decision rules, we utilized rule #4 (the scree plot rule) for the selection of the principal components. **We selected eight principal components because that explains 80% of the total variation in the data based on the visual inspection of the elbow in the scree plot.**

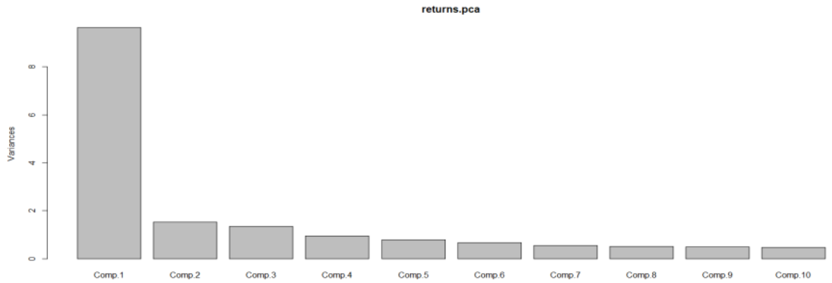


FIG 04: Default scree plot showing the amount of variance that each principal component contributes to the total, arranged from principal component 1 to principal component 20.

In FIG 04, we can confirm that the first principal component (the linear combination of the log-return variables) accounts for the highest sample variance in the data among all the linear combinations (principal components). It accounts for 48.18% of the total variance. The second principal component is defined as that linear combination of the log-return variables that accounts for a maximal proportion of the remaining variance subject to being uncorrelated with the first principal component. From Table 07, we can gather that the second principal component accounts for 7.69% of the remaining variance subject to being uncorrelated to the first principal component. The subsequent principal components are defined the same way.

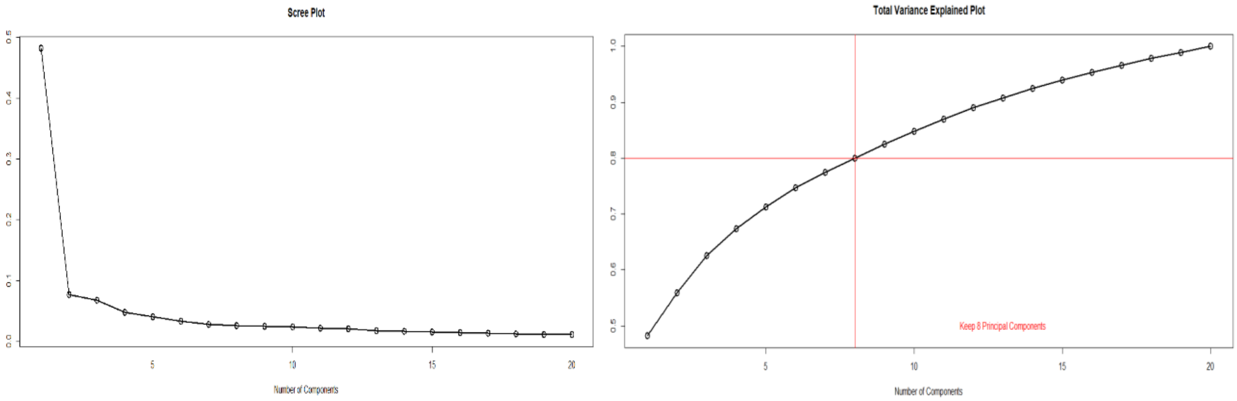


FIG 05: Two scree plots – a) plot of scree values (proportion of variance to total sum of the variance of all the data) for all components and b) line segment plot of cumulative variance for each principal component

From FIG 05 (a), we notice the elbow in the curve (a change of slope from "steep" to "shallow") occurs when the number of the principal components is 8. Also, from FIG 05 (b), it is evident based on the intersection of the y(cum. variance)=0.8 with the scree plot line that to be able to explain 80% of the variance in the variables, we need to keep the first 8 principal components. The same is also confirmed from the tabulation of the standard deviation, proportion of variance and cumulative proportion of each principal component in Table 07.

Principal Component	Standard deviation	Proportion of Variance	Cumulative Proportion
Comp.1	3.1042633	0.4818225	0.4818225
Comp.2	1.24037473	0.07692647	0.55874901
Comp.3	1.16079159	0.06737186	0.62612087
Comp.4	0.97348817	0.04738396	0.67350483
Comp.5	0.89191173	0.03977533	0.71328015
Comp.6	0.8163381	0.0333204	0.7466005
Comp.7	0.7472754	0.02792103	0.77452158
Comp.8	**0.71606462**	**0.02563743**	**0.800159**
Comp.9	0.70486968	0.02484206	0.82500107
Comp.10	0.68141987	0.02321665	0.84821772
Comp.11	0.65836107	0.02167196	0.86988968
Comp.12	0.6385577	0.0203878	0.8902775
Comp.13	0.592525	0.0175543	0.9078318
Comp.14	0.57888423	0.01675535	0.92458713
Comp.15	0.54494939	0.01484849	0.93943562
Comp.16	0.52563057	0.01381437	0.95324999
Comp.17	0.50927436	0.01296802	0.96621801
Comp.18	0.491994	0.0121029	0.9783209
Comp.19	0.4710181	0.0110929	0.9894138
Comp.20	0.46013436	0.01058618	1.000000

Table 07: Table shows the standard deviation, proportion of variance and cumulative proportion that each principal component explains

Code Snippet 05 in section 12.5 contains the R code for the principal component selection using the scree plots.

8. Principal components in predictive modeling

In this section, we present the results of fitting a linear regression model using the PCA scores by employing the first eight principal components (output of the previous section). The component scores or the PCA scores are the values of transformed variables based on the associated loadings value for each data point. We employed the predictive modeling framework to split the PCA scores into train and test data partitions based on 70/30 train/test (the basic form of cross-validation) split. We then computed the VIF values and the predictive accuracy metric (MAE) (using the train and test data partitions) for the fitted model.

Table 08 shows the PCA scores split into train and test partitions. The train/test split of the PCA scores is done using a uniform (0,1) random variable.

Data Partition	Number of records in the data partitions	Percentage of the records to the total records in PCA scores data set
Train	358	71.5%
Test	143	28.5%

Table 08: Table show record counts of the 70/30 training/test data partition of the PCA scores.

Next, we fitted a linear regression model using the first eight principal components to predict the log-return value of VV index fund using the train PCA scores partition. The resulting fitted regression equation is

VV = 7.620e-04+ 2.202e-03 * Comp.1 - 3.950e-04 * Comp.2 - 4.786e-04 * Comp.3 +2.711e-05 * Comp.4 + 2.567e-04 * Comp.5 + 2.151e-04 * Comp.6 - 3.484e-04 * Comp.7 -3.619e-04 * Comp.8

Table 09 shows the regression summary table for the fitted model (we refer to it as pca.lm)

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.62E-04	1.41E-04	5.409	1.18e-07 ***
Comp.1	2.20E-03	4.53E-05	48.615	< 2e-16 ***
Comp.2	-3.95E-04	1.12E-04	-3.539	0.000457 ***
Comp.3	-4.79E-04	1.29E-04	-3.718	0.000234 ***
Comp.4	2.71E-05	1.42E-04	0.191	0.848949
Comp.5	2.57E-04	1.57E-04	1.638	0.102272
Comp.6	2.15E-04	1.69E-04	1.271	0.204677
Comp.7	-3.48E-04	1.97E-04	-1.77	0.077617 .
Comp.8	-3.62E-04	1.92E-04	-1.882	0.060692 .

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002658 on 349 degrees of freedom
Multiple R-squared:  0.8746,    Adjusted R-squared:  0.8717
F-statistic: 304.2 on 8 and 349 DF,  p-value: < 2.2e-16
Residuals:
        Min        1Q    Median        3Q       Max
-0.0126119 -0.0015016  0.0000017  0.0015389  0.0102312
```

Table 09: Regression summary table for the pca.lm model lm(formula = VV ~ Comp.1 + Comp.2 + Comp.3 + Comp.4 + Comp.5 + Comp.6 + Comp.7 + Comp.8, data = train.scores)

From Table 09, we can determine that at a level of significance of 0.001, the overall regression is strongly statistically significant with a p-value < 2.2e-16 (with a F-statistic value of 304.2). Therefore, at a level of significance of 0.001, we can reject the overall F-test (joint) null hypothesis that all the predictor variables have a zero coefficient and have no effect on the regression. Regarding the individual regression coefficients, only the intercept, Comp.1, Comp.2, and Comp.3 are strongly statistically significant at a level of significance of 0.001 (meaning at level of significance of 0.001, we can reject the null hypothesis that the individual regression coefficient for these components is zero). The coefficients for Comp.7 and Comp.8 are statistically significant at the level of significance 0.1. All the rest of the coefficients, namely Comp.4, Comp.5, and Comp.6, are NOT statistically significant. Interestingly only the first few principal components used in the regression are significant.

Also, from the Table 09, we can determine that the coefficient of determination (adjusted-R-squared) is 87.17% indicating that 87.17 percent of variation in log-return of VV can be explained variation in the principal components.

We then computed the VIF values for the principal components utilized in the fitting of the model. Table 10 shows the VIF values for the first 8 principal components.

Variable (Principal Component)	VIF value
Comp.1	1.007069
Comp.2	1.009490
Comp.3	1.007559
Comp.4	1.018314
Comp.5	1.009776
Comp.6	1.008544
Comp.7	1.010238
Comp.8	1.007752

Table 10: VIF values for all the first 8 principal components used in the fitting of the pca.lm model

None of the VIF values are greater than 3. All the values are approximately 1 (~ 1). Multicollinearity is not an issue with the principal components. The VIF values associated with every predictor variable in any principal components regression model should be one. This is because in the formula to compute the VIF value,

$$VIF_j = \frac{1}{(1 - Rj^2)}$$

In the absence of any linear relationship between the predictor variables (the principal components in the case of pca.lm), Rj^2 would be zero. As a result, VIF_j would be one for all the predictor variables in any principal components regression models.

We next computed the predictive accuracy metric, MAE (Mean Absolute Error) using the train and test partitions of the PCA scores data. Table 11 shows the in-sample (train) and out-of-sample (test) MAE values. The MAE value computed for the test data is only slightly smaller than the MAE value computed using the train data.

Model	MAE value using the train data	MAE value using the test data
pca.lm	0.001973056	0.002255475

Table 11: MAE values for pca.lm using train and test data partitions

Code Snippet 06 in section 12.6 contains the R code for fitting the linear regression model using the first eight principal components to predict the log-return VV index fund using the PCA scores data.

9. Model comparison

In this section, we present the results of the comparison of the models pca.lm (the model from previous section that is fitted with the first eight principal components to predict the log-return vv index fund value) and model.1 (the small model which is fitted with the log-return stock variables for GS, DD, DOW, HON, HUN, JPM, KO, MMM and XOM to predict the log-return VV stock value) and model.2 (the full model which is fitted with all the log-return stock variables for the prediction of the log-return VV stock value). Both the model.1 (small model) and model.2 (full model) were fitted in section 5 using the log-return data as part of the exploratory data analysis to examine the VIF values. However, in this section, we present the results of the model.1 (small model) and model.2 (full model) which were re-fitted using the train data partition. We employed the predictive modeling framework technique to split the returns data into train/test data partitions based on the 70/30 train/test split.

The returns data is split using the same random (0,1) sampling that was employed for the splitting of the PCA scores data in section 8. As a result, the number of records in the train and the test data partitions shown in Table 12 are exactly same as for the PCA scores data split shown in section 8.

Data Partition	Number of records in the data partitions	Percentage of records to the total records in returns data set
Train	358	71.5%
Test	143	28.5%

Table 12: Table show record counts of the 70/30 training/test data partition of the returns dataset.

The re-fitting for the small model (model.1) using returns train partition results in the following equation:

VV = 0.0001291+ 0.0897822 * GS + 0.0273590 * DD + 0.0373641 * DOW + 0.1527107 * HON + 0.0334689 * HUN + 0.0487792 * JPM + 0.1407796 * KO + 0.1276387 * MMM + 0.1231533 * XOM

Based on the summary table in Table 13, we can conclude at the level of significance of 0.001, we reject that the joint null hypothesis that all the coefficients have no effect on the overall regression. This is based on the p-value (with a corresponding F-statistic value of 230.5) of less than 2.2e-16. Also, it can be gleaned from Table 13 is that the individual regression coefficients for GS, HON, HUN, KO, MMM, and XOM are strongly significant at the level of significance of 0.001 meaning we can reject the null hypothesis for these individual regression coefficients to be zero at the level of significance of 0.001. The regression coefficients for the log return variables DOW and JPM are only significant at a level of significance of 0.01. The intercept and coefficient for log-return variable DD are NOT significant.

The coefficient of determination for the small model is 85.27% indicating that 85.27 percent of the variation in the log-return of VV index fund can be explained by variations in the predictor variables.

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0001291	0.0001521	0.849	0.396656
GS	0.0897822	0.0158585	5.661	3.14e-08 ***
DD	0.027359	0.0190828	1.434	0.152557
DOW	0.0373641	0.0127952	2.92	0.003726 **
HON	0.1527107	0.0188163	8.116	8.41e-15 ***
HUN	0.0334689	0.0086227	3.881	0.000124 ***
JPM	0.0487792	0.0150467	3.242	0.001302 **
KO	0.1407796	0.0205399	6.854	3.28e-11 ***
MMM	0.1276387	0.0262976	4.854	1.83e-06 ***
XOM	0.1231533	0.0239697	5.138	4.64e-07 ***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002849 on 348 degrees of freedom
Multiple R-squared:  0.8564,    Adjusted R-squared:  0.8527
F-statistic: 230.5 on 9 and 348 DF,  p-value: < 2.2e-16
Residuals:
       Min        1Q     Median        3Q       Max
-0.0137752 -0.0015139  0.0000135  0.0016244  0.0086977
```

Table 13: Regression summary table after re-fitting the model.1 (Small) model lm (formula = VV ~ GS + DD + DOW + HON + HUN + JPM + KO + MMM + XOM, data = train.returns)

The re-fitting for the full model using returns train data paration results in the following equation:

$$\mathbf{VV = 0.0001421 + 0.0366395 * BAC + 0.0505577* GS + 0.0099939 * JPM + 0.0798350 * WFC + 0.0248740 * BHI + 0.0597736 * CVX + 0.0034018 * DD + 0.0343876 * DOW + 0.0407020 * DPS + 0.0165769 * HAL - 0.0148023 * HES + 0.1192409 * HON + 0.0240021 * HUN + 0.1019110 * KO + 0.1035769 * MMM + 0.0118111 * MPC + 0.0271242 * PEP + 0.0352489 * SLB + 0.0408361 * XOM}$$

Using Table 14, we can conclude that at a level of significance of 0.001, we can reject the joint null hypothesis that all the coefficients have no effect on the overall regression. This is based on the p-value (F-statistic is 141.6) of less than 2.2e-16.

For the individual coefficient tests, the variables for the log returns of the stocks BAC, WFC, HON, KO and MMM are strongly statistically significant at the level of significance of 0.001. The coefficients for GS, DOW, and HUN are significant at only at a level of significance of 0.01. The coefficients for CVX, DPS and SLB are significant at level of significance 0.05. Coefficient for BHI is only significant at a level of significance of 0.1. The rest of the coefficients for the variables JPM, DD, HAL, HES, MPC, PEP, XOM and the intercept are not significant.

Also, from the summary table Table 14, we can determine that the coefficient of determination for the full model (model.2) is 88.21% indicating 88.21 percent of the variation in the log return of VV can be explained by the variations in the predictor variables used in the full model.

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0001421	0.0001372	1.036	0.301068
BAC	0.0366395	0.0108814	3.367	0.000847 ***
GS	0.0505577	0.0152744	3.31	0.001034 **
JPM	0.0099939	0.0149494	0.669	0.50426
WFC	0.079835	0.0195645	4.081	5.61e-05 ***
BHI	0.024874	0.0134781	1.846	0.065837 .
CVX	0.0597736	0.0243308	2.457	0.014524 *
DD	0.0034018	0.0174049	0.195	0.845157
DOW	0.0343876	0.0116107	2.962	0.003276 **
DPS	0.040702	0.0166255	2.448	0.014866 *
HAL	0.0165769	0.0139689	1.187	0.236179
HES	-0.0148023	0.0102847	-1.439	0.151002
HON	0.1192409	0.0175181	6.807	4.56e-11 ***
HUN	0.0240021	0.0080033	2.999	0.002909 **
KO	0.101911	0.02137	4.769	2.76e-06 ***
MMM	0.1035769	0.0239288	4.329	1.98e-05 ***
MPC	0.0118111	0.0080854	1.461	0.145002
PEP	0.0271242	0.02322	1.168	0.243575
SLB	0.0352489	0.0161827	2.178	0.030083 *
XOM	0.0408361	0.0252549	1.617	0.10682

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002548 on 338 degrees of freedom
Multiple R-squared:  0.8884,    Adjusted R-squared:  0.8821
F-statistic: 141.6 on 19 and 338 DF,  p-value: < 2.2e-16
Residuals:
        Min       1Q    Median       3Q      Max
-0.013584 -0.001442  0.000094  0.001551  0.007602
```

Table 14: Regression summary table after re-fitting the model.2 (full) model lm(formula = VV ~ BAC + GS + JPM + WFC + BHI + CVX + DD + DOW + DPS + HAL + HES + HON + HUN + KO + MMM + MPC + PEP + SLB + XOM, data = train.returns)

Next, we computed the predictive accuracy metric MAE for model.1 and model.2 using the in-sample and out-of-sample data. MAE measures the average magnitude of the errors in the predictions, without considering the direction. Table 15 shows the MAE values for model.1 and model.2 next to pca.lm for comparison. Smaller MAE values are preferred. The training MAE errors are lower than the test MAE errors indicating overfitting to the training data.

Model	MAE value using the train data	MAE value using the test data
pca.lm	0.001973056	**0.002255475**
Model.1 (small model)	0.002124989	0.0023513
Model.2 (full model)	**0.001900968**	0.002266901

Table 15: MAE values for the three models – pca.lm, model.1, model.2 using in-sample and out-of-sample data

From the table, we can gather that the model with the smallest MAE computed using the train data is Model.2 (full model). However, the model with the smallest MAE value in the case where the test data partition was used for MAE computation is pca.lm. **Since the out-of-sample MAE value are used for accessing the predictive performance, we pick pca.lm as the best model among the three models.**

Code Snippet 07 in section 12.7 contains the R code for model comparison using pca.lm, model.1 (small model), and model.2 (full model)

10. Automatic variable selection for the principal components to keep

In this section, we present the results of employing the automatic variable selection method to select the principal components to keep. We used backward variable selection method for this. We utilized the stepAIC function in R with "backward" direction and a starter regression equation containing a full model with all the principal components to predict the log-return VV variable. The train partition for the PCA scores is used. The stepAIC function performed 9 iterations based on the AIC value. The backward variable selection method selected a regression equation with 14 terms (13 principal components and an intercept term).

The fitted regression equation selected by the backward variable selection (we refer to it as backward.lm) is

VV = 7.977e-04 + 2.200e-03 * Comp1. -4.133e-04 * Comp.2 -5.013e-04 * Comp.3 + 2.294e-04 * Comp.5 + 2.257e-04 * Comp.6 -3.709e-04 * Comp.7 -3.565e-04 * Comp.8 + 5.992e-04 * Comp.9 -4.109e-04 * Comp.10 + 6.843e-04 * Comp.11 + 3.443e-04 * Comp.12 -5.057e-04 * Comp.14 -4.561e-04 * Comp.20

Compared to the first 8 principal components used in the pca.lm model, the backward variable selection suggests 13 principal components. Among the first 8, the backward variable selection does not suggest Comp.4 and Comp.5.

On the principal components picked out by the stepAIC function using the backward variable selection, we computed the VIF values (Variance Inflation Factors) to determine if any of them are collinear. Typically, VIF values greater than 3 indicate the presence of multicollinearity among the predictor variables. From Table 16, we can gather the VIF values are approximately ~ 1.

Principal Component	VIF value
Comp.14	1.019139
Comp.11	1.017025
Comp.7	1.014280
Comp.3	1.013893
Comp.10	1.013552
Comp.6	1.012671
Comp.2	1.011968
Comp.8	1.010650
Comp.9	1.008218
Comp.1	1.007825
Comp.20	1.007731
Comp.5	1.006564
Comp.12	1.004756

Table 16: VIF values for the principal components picked out the backward variable selection model

From the Table 17 below, we can conclude that at a level of significance of 0.001, the overall regression is strongly statistically significant with a p-value < 2.2e-16 (with a F-statistic value of 208.3). So, at a level of significance of 0.001, we can reject the null joint hypothesis that all the predictor variables have a zero coefficient and have no effect on the regression.
Regarding the individual regression coefficients, the coefficients for the principal components Comp.1, Comp.2, Comp.3, Comp.11, and the intercept are strongly statistically significant at the level of significance 0.001. The coefficient for Comp.9 is statistically significant at the level of significance 0.01.

The coefficients for the principal components Comp.7, Comp.10, and Comp.14 are statistically significant at a level of significance 0.05. The coefficients for the components Comp.8 and Comp.12 are significant at a level of significance of 0.1. The coefficients for Comp.5, Comp.6, and Comp.20 are NOT significant. The adjusted-r-squared value is 0.883 indicating 88.3% of the variation in the log-return of VV is explained by the variations in the principal components.

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.98E-04	1.35E-04	5.929	7.42e-09 ***
Comp.1	2.20E-03	4.33E-05	50.861	< 2e-16 ***
Comp.2	-4.13E-04	1.07E-04	-3.873	0.000128 ***
Comp.3	-5.01E-04	1.23E-04	-4.065	5.95e-05 ***
Comp.5	2.29E-04	1.49E-04	1.536	0.125372
Comp.6	2.26E-04	1.62E-04	1.394	0.164182
Comp.7	-3.71E-04	1.88E-04	-1.97	0.049644 *
Comp.8	-3.57E-04	1.84E-04	-1.939	0.053286 .
Comp.9	5.99E-04	1.86E-04	3.218	0.001412 **
Comp.10	-4.11E-04	1.93E-04	-2.132	0.033754 *
Comp.11	6.84E-04	2.05E-04	3.34	0.000931 ***
Comp.12	3.44E-04	2.02E-04	1.706	0.088920 .
Comp.14	-5.06E-04	2.29E-04	-2.205	0.028112 *
Comp.20	-4.56E-04	2.95E-04	-1.545	0.123228

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002538 on 344 degrees of freedom
Multiple R-squared:  0.8873,    Adjusted R-squared:  0.883
F-statistic: 208.3 on 13 and 344 DF,  p-value: < 2.2e-16
Residuals:
       Min        1Q    Median        3Q       Max
-0.0133766 -0.0014590  0.0001723  0.0015815  0.0078665
```

Table 17: Regression output for the estimated model using the backward variable selection (backward.lm)

Next, we computed the predictive accuracy metric MAE for the model estimated using the backward variable selection. This is done using both the in-sample (train) and out-of-sample (test) data partitions. Table 18 shows the MAE values for the four models (pca.lm, model.1 (small model), model.2 (full model), and backward.lm) using the train and test data partitions.

Model	MAE value using the train data	MAE value using the test data
pca.lm	0.001973056	**0.002255475**
Model.1 (small model)	0.002124989	0.0023513
Model.2 (full model)	**0.001900968**	0.002266901
backward.lm	0.001915101	0.002266125

Table 18: MAE values for the four models – pca.lm, model.1 (small model), model.2 (full model), backward.lm using in-sample and out-of-sample data

From Table 18, we can determine that Model.2 (the full model) has the lowest MAE value among all the four models using the in-sample data. Among the MAE values computed using the out-of-sample data partition, the pca.lm (the linear regression model that is fitted with the first eight principal components) has the lowest MAE value.

Based on this comparison, pca.lm (the linear model to predict the log-return value of Vanguard (VV) index fund using the first eight principal components) is the best model among all the fitted models.

Code Snippet 08 in section 12.8 contains the R code for automatic variable selection for the principal components to keep and the model comparison results for the four models - pca.lm, model.1 (small model), model.2 (full model), backward.lm using the in-sample and the out-of-sample data.

11.Summary

We began the analysis by preparing the data for the principal component analysis. We applied log-return transformations on the raw stock data. We also sorted the data based on the date (after the date using R Date format).

Following that, we performed exploratory data analysis using statistical graphs and data visualization tools to identify correlation in the data. After that, more exploratory data analysis is performed using two naïve models (a small model and a full model). VIF values were computed on the predicted variables to detect multicollinearity in the data. Since we had a couple of variables with VIF values greater than three, multicollinearity is an issue in the data.

We then performed the principal component analysis on the log-returns data. We used scree plots to select the first eight principal components as they collectively explain 80% of the variation in the data. This corresponding to an unsupervised decision rule.

Next, we fitted a linear regression model to predict the log-return value of the Vanguard large-cap index fund using the first eight principal components. The component scores data used for fitting the model was split into train/test partitions using the predictive modeling framework. This fitted model is then compared with the two naïve models (a small model and a full model) which were re-fitted with the train data (the returns data is split based on the predictive modeling framework using the same random sampling as applied to the component score data). The predictive accuracy metric, MAE, was used for the comparison. The full model has the lowest MAE value among the in-sample computations. But the linear model fitted with the first 8 principal components has the lowest MAE value among the out-of-sample computations.

Lastly, we used the automatic variable selection (backward variable selection method) as a supervised approach for selecting the number of principal components using the component scores data. The resultant model using the backward variable selection method was compared with the linear model fitted the PCA's first eight components and the naïve models using the predictive accuracy metric, MAE. **The linear model fitted with PCA's first eight components was the model with the lowest MAE value (computed using test data) among the four models and hence the best performing model.**

12.Code

12.1. Data preparation

```
# Read in csv file for stock portfolio data
my.path <- 'C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week6_7\\Assignment_06\\Code\\';
my.data <- read.csv(paste(my.path,'stock_portfolio.csv',sep=''),header=TRUE);

# View the first few rows and the structure of the data
head(my.data)
str(my.data)

# Note Date is a string of dd-Mon-yy in R this is '%d-%B-%y';
my.data$RDate <- as.Date(my.data$Date,'%d-%B-%y');
# sort the data based on Date
sorted.df <- my.data[order(my.data$RDate),];
head(sorted.df)

# compute the log return value of the stock values
AA <- log(sorted.df$AA[-1]/sorted.df$AA[-dim(sorted.df)[1]]);
# Manually check the first entry: log(9.45/9.23)
# Type cast the array as a data frame;
returns.df <- as.data.frame(AA);
returns.df$BAC <- log(sorted.df$BAC[-1]/sorted.df$BAC[-dim(sorted.df)[1]]);
returns.df$BHI <- log(sorted.df$BHI[-1]/sorted.df$BHI[-dim(sorted.df)[1]]);
returns.df$CVX <- log(sorted.df$CVX[-1]/sorted.df$CVX[-dim(sorted.df)[1]]);
returns.df$DD <- log(sorted.df$DD[-1]/sorted.df$DD[-dim(sorted.df)[1]]);
returns.df$DOW <- log(sorted.df$DOW[-1]/sorted.df$DOW[-dim(sorted.df)[1]]);
returns.df$DPS <- log(sorted.df$DPS[-1]/sorted.df$DPS[-dim(sorted.df)[1]]);
returns.df$GS <- log(sorted.df$GS[-1]/sorted.df$GS[-dim(sorted.df)[1]]);
returns.df$HAL <- log(sorted.df$HAL[-1]/sorted.df$HAL[-dim(sorted.df)[1]]);
returns.df$HES <- log(sorted.df$HES[-1]/sorted.df$HES[-dim(sorted.df)[1]]);
returns.df$HON <- log(sorted.df$HON[-1]/sorted.df$HON[-dim(sorted.df)[1]]);
returns.df$HUN <- log(sorted.df$HUN[-1]/sorted.df$HUN[-dim(sorted.df)[1]]);
returns.df$JPM <- log(sorted.df$JPM[-1]/sorted.df$JPM[-dim(sorted.df)[1]]);
returns.df$KO <- log(sorted.df$KO[-1]/sorted.df$KO[-dim(sorted.df)[1]]);
returns.df$MMM <- log(sorted.df$MMM[-1]/sorted.df$MMM[-dim(sorted.df)[1]]);
returns.df$MPC <- log(sorted.df$MPC[-1]/sorted.df$MPC[-dim(sorted.df)[1]]);
returns.df$PEP <- log(sorted.df$PEP[-1]/sorted.df$PEP[-dim(sorted.df)[1]]);
returns.df$SLB <- log(sorted.df$SLB[-1]/sorted.df$SLB[-dim(sorted.df)[1]]);
returns.df$WFC <- log(sorted.df$WFC[-1]/sorted.df$WFC[-dim(sorted.df)[1]]);
returns.df$XOM <- log(sorted.df$XOM[-1]/sorted.df$XOM[-dim(sorted.df)[1]]);
returns.df$VV <- log(sorted.df$VV[-1]/sorted.df$VV[-dim(sorted.df)[1]]);
```

Code Snippet 01: R code for the data preparation

12.2. Exploratory data analysis using statistical graphs and data visualization tools

```
# Compute correlation matrix for returns;
returns.cor <- cor(returns.df)
returns.cor[,c('VV')]
# Barplot the last column to visualize magnitude of correlations;
barplot(returns.cor[1:20,c('VV')],las=2,ylim=c(0,1.0))
title('Correlations with VV')

# load the corrplot package
library(corrplot)
# Make correlation plot for returns;
corrplot(returns.cor,method="number",number.cex=0.75)
```

Code Snippet 02: R code for the Exploratory Data Analysis (EDA) using statistical graphs and data visualization plots.

12.3. Exploratory data analysis using models

```
# load the car package
library(car)
# Fit the small model
model.1 <- lm(VV ~ GS+DD+DOW+HON+HUN+JPM+KO+MMM+XOM, data=returns.df)
summary(model.1)
sort(vif(model.1), decreasing=TRUE)

# Fit the full model
model.2 <- lm(VV ~
AA+BAC+GS+JPM+WFC+BHI+CVX+DD+DOW+DPS+HAL+HES+HON+HUN+KO+MMM+MPC+PEP+SLB+XOM,data=returns.df)
summary(model.2)
sort(vif(model.2),decreasing=TRUE)
```

Code Snippet 03: R code for the Exploratory Data Analysis (EDA) using models

12.4. Variable transformation using the principal component analysis

```
# compute the principal components using the log-return data
returns.pca <- princomp(x=returns.df[,-21],cor=TRUE)

# See the output components returned by princomp();
names(returns.pca)
pc.1 <- returns.pca$loadings[,1];
pc.2 <- returns.pca$loadings[,2];
names(pc.1)

# create a plot of the principal components pc.1 vs pc.2
plot(-10,10,type='p',xlim=c(0.12,0.27),ylim=c(-0.60,0.25),xlab='PC 1',ylab='PC 2')
text(pc.1,pc.2,labels=names(pc.1),cex=1.25,col=c(7,2,3,4,5,5,6,2,3,4,1,5,2,6,1,4,6,3,2,4,9))
title("plot of pc.1 vs pc.2")
```

Code Snippet 04: R code for the principal component analysis using the log-return data

12.5. Principal component selection

```
# Plot the default scree plot;
plot(returns.pca)

scree.values <- (returns.pca$sdev^2)/sum(returns.pca$sdev^2);
plot(scree.values,xlab='Number of Components',ylab='',type='l',lwd=2)
points(scree.values,lwd=2,cex=1.5)
title('Scree Plot')

# Make Proportion of Variance Explained
variance.values <- cumsum(returns.pca$sdev^2)/sum(returns.pca$sdev^2);
plot(variance.values,xlab='Number of Components',ylab='',type='l',lwd=2)
points(variance.values,lwd=2,cex=1.5)
abline(h=0.8,lwd=1.5,col='red')
abline(v=8,lwd=1.5,col='red')
text(13,0.5,'Keep 8 Principal Components',col='red')
title('Total Variance Explained Plot')
```

Code Snippet 05: R code for the principal component selection using scree plots

12.6. Principal components in predictive modeling

```
# Create the data frame of PCA predictor variables;
return.scores <- as.data.frame(returns.pca$scores);
return.scores$VV <- returns.df$VV;
set.seed(123)
return.scores$u <- runif(n=dim(return.scores)[1],min=0,max=1);
head(return.scores)

# Split the data set into train and test data sets;
train.scores <- subset(return.scores,u<0.70);
test.scores <- subset(return.scores,u>=0.70);
dim(train.scores)
dim(test.scores)
dim(train.scores)+dim(test.scores)
dim(return.scores)

# Fit a linear regression model using the first 8 principal components;
pca1.lm <- lm(VV ~ Comp.1+Comp.2+Comp.3+Comp.4+Comp.5+Comp.6+Comp.7+Comp.8,data=train.scores);
summary(pca1.lm)

# Compute the Mean Absolute Error on the training sample;
pca1.mae.train <- mean(abs(train.scores$VV-pca1.lm$fitted.values));
vif(pca1.lm)

# Score the model out-of-sample and compute MAE;
pca1.test <- predict(pca1.lm,newdata=test.scores);
pca1.mae.test <- mean(abs(test.scores$VV-pca1.test));
```

Code Snippet 06: R code for fitting the linear regression model using the first eight principal components to predict the log-return VV index fund using PCA scores data.

12.7. Model comparison

```
# Let's compare the PCA regression model with a 'raw' regression model;
# Create a train/test split of the returns data set to match the scores data set;
returns.df$u <- return.scores$u;
train.returns <- subset(returns.df,u<0.70);
test.returns <- subset(returns.df,u>=0.70);
dim(train.returns)
dim(test.returns)
dim(train.returns)+dim(test.returns)
dim(returns.df)

# Fit model.1 on train data set and score on test data;
model.1 <- lm(VV ~ GS+DD+DOW+HON+HUN+JPM+KO+MMM+XOM, data=train.returns)
summary(model.1)
model1.mae.train <- mean(abs(train.returns$VV-model.1$fitted.values));
model1.test <- predict(model.1,newdata=test.returns);
model1.mae.test <- mean(abs(test.returns$VV-model1.test));

# Fit model.2 on train data set and score on test data;
model.2 <- lm(VV ~
BAC+GS+JPM+WFC+BHI+CVX+DD+DOW+DPS+HAL+HES+HON+HUN+KO+MMM+MPC+PEP+SLB+XOM,data=train.returns)
summary(model.2)
model2.mae.train <- mean(abs(train.returns$VV-model.2$fitted.values));
model2.test <- predict(model.2,newdata=test.returns);
model2.mae.test <- mean(abs(test.returns$VV-model2.test));
```

Code Snippet 07: R code for model comparison using pca.lm, model.1 (small model), and model.2 (full model)

12.8. Automatic variable selection to pick the principal components

```
# full model for use in the backward variable selection
full.lm <- lm(VV ~ ., data=train.scores);
summary(full.lm)

# load the MASS library
library(MASS)

# Automatic variable selection of principal component selection
backward.lm <- stepAIC(full.lm,direction=c('backward'))
summary(backward.lm)

# MAE values using train and test partitions
backward.mae.train <- mean(abs(train.scores$VV-backward.lm$fitted.values))
backward.test <- predict(backward.lm,newdata=test.scores)
backward.mae.test <- mean(abs(test.scores$VV-backward.test))

# VIF values for the backward.lm
vif(backward.lm)
```

Code Snippet 08: R code for automatic variable selection to pick the principal components