Assignment 6: Principle Components in Predictive Modeling

Sri Seshadri

7/27/2017

1. Introduction

WRITE UP GOES HERE

2. Exploratory analysis

Since there are number of variables in the data, we start by quantifying the associations amongst them by plotting the correlations. The index VV is of interest and it would be useful to plot the correlation of VV with other variables. Figure 1 shows the correlations of VV with other stock indices.

2.1 Statistical graphic Vs Data visualization

While figure 1 is useful for understanding the VV's correlations with other stock indices, it would be useful to gain insights into how other stock indicies are correlated amongst themselves and how the correlations compare relatively. From figure 2, it can be noted that higher correlations are on the base of the right triangles formed by the matrx diagonal i.e. indices SLB, WFC, XOM and VV are relatively highly correlated with other indices. Such insights are possible only with visualizations such as this.

It can be seen that indices DPS, MPC and PEP are not significantly correlated with other indices. They are likely to have low Variance Inflation Factors (VIF), as mentioned above GS, XOM, SLB and VV are likely to have higher VIF due to their high correlations with other indices.

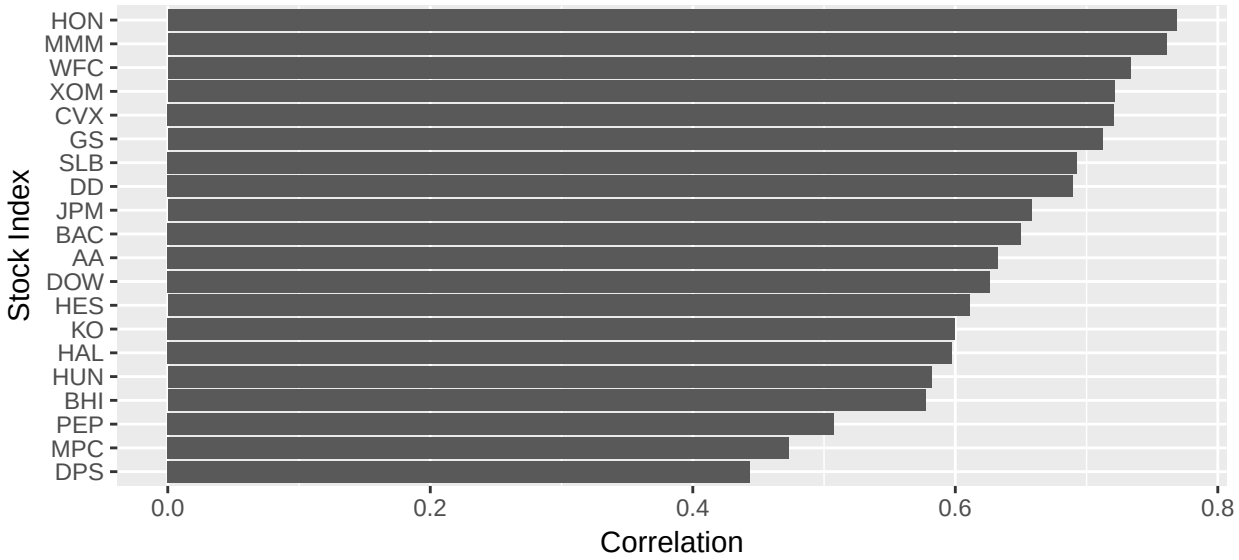

Figure 1: Correlations with VV

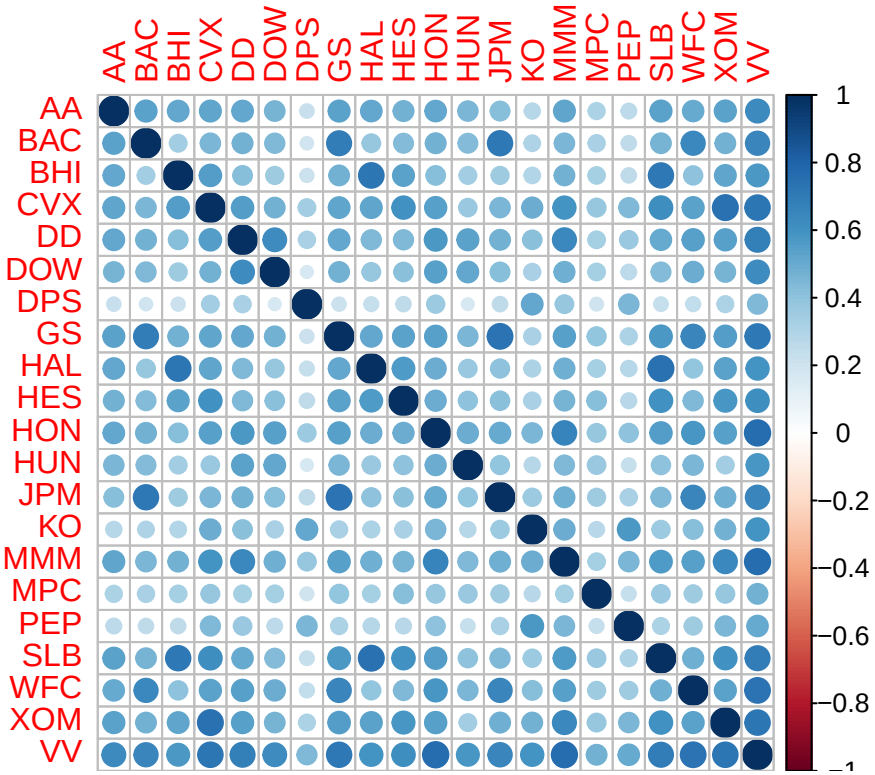

Figure 2: Correlation plot

2.2 Variance Inflation Factor (VIF)

In the previous section, it was seen that GS, XOM, and SLB were highly correlated with other predictors and they were suspected to have relatively high VIF. One way of assessing the VIF is to fit regression models and assessing VIF. Two models; an arbitary model and a full model. The VIFs are assesed for the predictor variables in the model and is shown below in table 1. Table 1 shows the top 5 VIFs for each model. It is seen that variance inflation factors are less than the thumb rule of 10, evidence of multicolinearity not to be a concern. It would be interesting to see how Principle Component Analysis (PCA) would work on this data.

```
##
## Call:
## lm(formula = VV ~ GS + DD + DOW + HON + HUN + JPM + KO + MMM +
##     XOM, data = returns)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0139179 -0.0016005 -0.0000926  0.0016690  0.0172703
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0001008  0.0001331   0.757 0.449290
## 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
```

```
##
## Call:
## lm(formula = VV ~ ., data = returns)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0139266 -0.0015542  0.0000313  0.0015042  0.0140759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.953e-05  1.213e-04   0.820 0.412433
## AA           1.538e-02  1.040e-02   1.479 0.139894
## BAC          2.723e-02  9.697e-03   2.808 0.005188 **
## BHI          1.604e-02  1.161e-02   1.382 0.167752
## CVX          5.742e-02  2.068e-02   2.776 0.005720 **
## DD           1.003e-02  1.625e-02   0.618 0.537144
## DOW          3.600e-02  1.069e-02   3.366 0.000824 ***
## DPS          5.659e-02  1.493e-02   3.790 0.000170 ***
```

Table 1: Top 5 VIF by model

Model	Predictors	VIF
Arbitary model	GS	2.705795
	MMM	2.590177
	DD	2.368257
	JPM	2.324600
	HON	2.261397
Full Model	SLB	3.258982
	GS	3.197811
	XOM	2.949826
	CVX	2.920187
	HAL	2.919924

```
## GS           3.434e-02  1.358e-02   2.529 0.011766 *
## HAL         -1.976e-03  1.210e-02  -0.163 0.870344
## HES          4.393e-03  9.688e-03   0.453 0.650432
## HON          1.071e-01  1.608e-02   6.658 7.62e-11 ***
## HUN          2.867e-02  7.222e-03   3.969 8.31e-05 ***
## JPM          2.224e-02  1.329e-02   1.674 0.094849 .
## KO           9.425e-02  1.847e-02   5.103 4.82e-07 ***
## MMM          1.093e-01  2.202e-02   4.964 9.63e-07 ***
## MPC          1.079e-02  7.024e-03   1.536 0.125134
## PEP          2.092e-02  2.034e-02   1.028 0.304293
## SLB          4.851e-02  1.453e-02   3.339 0.000905 ***
## WFC          7.738e-02  1.580e-02   4.897 1.33e-06 ***
## XOM          5.797e-02  2.301e-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
```

3. Principle Component Analysis (PCA)

The tickers from AA through XOM (not including VV) are transformed using PCA. The scatter plot of loadings (loadings are the weights of the variables that contribute to the component - the columns of the loading matrix correspond to eigen vectors) of the first two principle components is shown in Figure 3. It is seen that the tickers belonging to soft drinks industry are grouped together. They also have a higher weightage in explaining the variation in the data. ANY SURPRISES? It will be interesting to see how PCA reduced the data in terms of variation.

3.1 Scree plot

Scree plots are useful in identifying the principle components that contribute to explaining the variation in the variables. Figures 4, 5 and 6 show different variations of the scree plot. Figure 4 is the default version of scree plot in R. The scree plot looks to be truncated to 10 components.It is seen that component 1 explains the majority of the variance.

It would be useful to see the proportion of variance explained by the components. Figure 5 shows the proportion of the variance that the principle components are able to explain. It is seen that the contribution

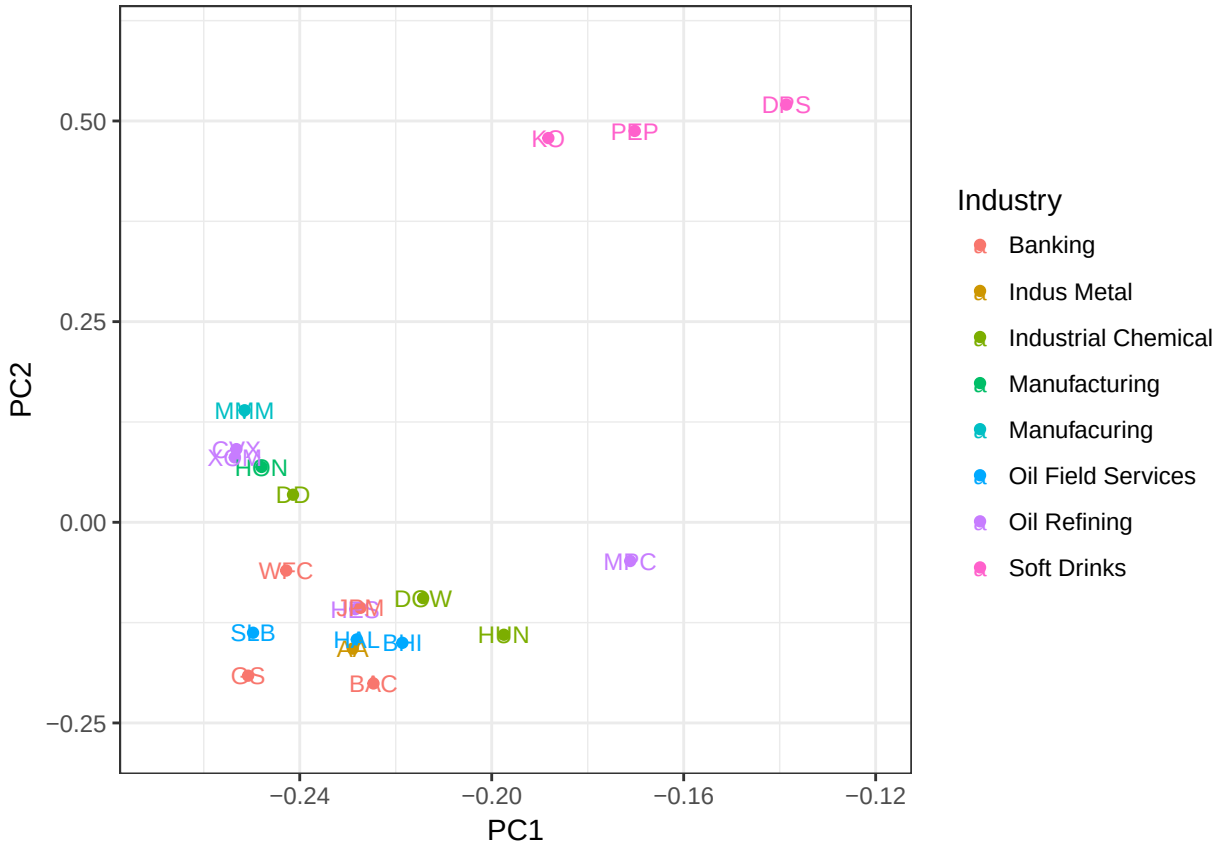

Figure 3: Variable loadings of first 2 principle components

of principle components to explanation of variance plateaus at component number 8. Additional components beyond 8 are not a significant addition to explaining the variance. As a rule of thumb, principle components that explain 80% (or 70% to 90%) of the total variance is chosen for data reduction. It will be useful to plot the cumulative variance of the principle components. Figure 6 shows the first 8 components explain 80% of the total variance in the data.

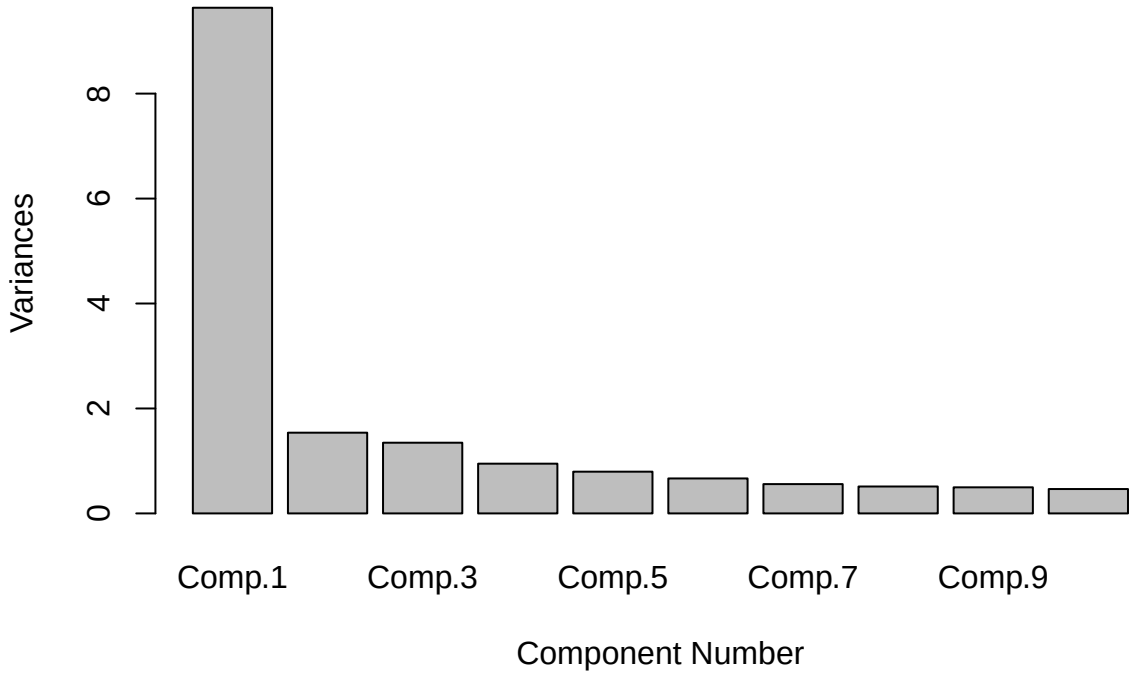

Figure 4: Scree Plot - Not so good looking

With the data reduced to eight principle components, the eight components are to be used as predictors for modelling. The next section discusses the model

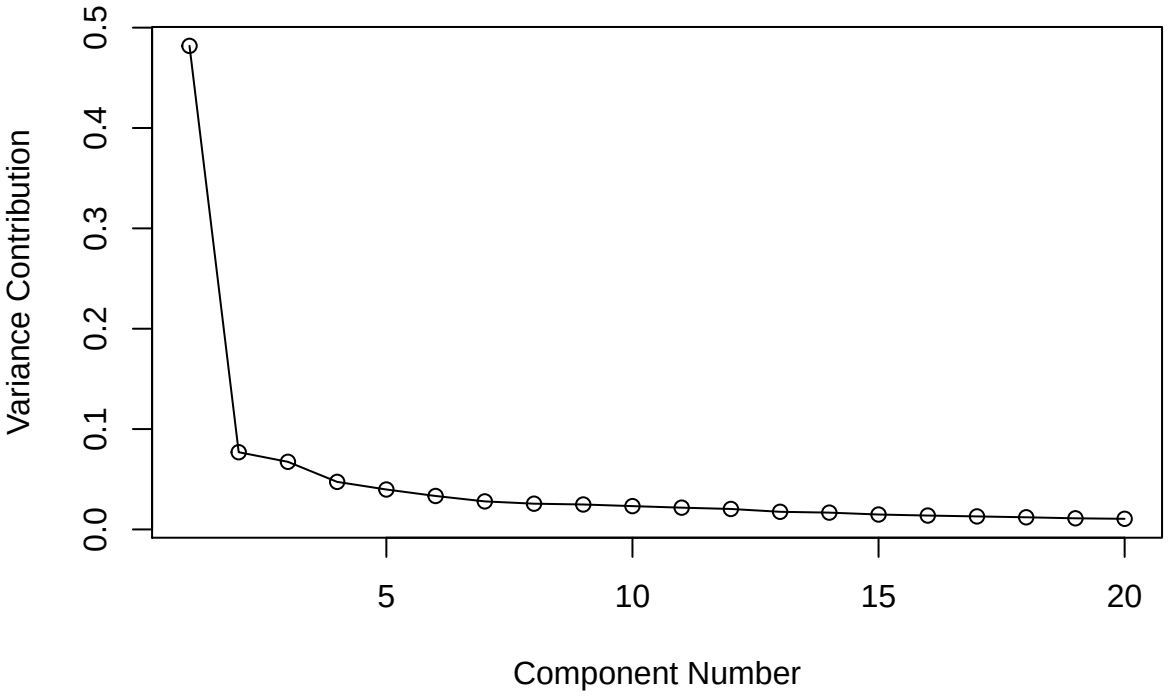

Figure 5: Scree plot - Variance explained

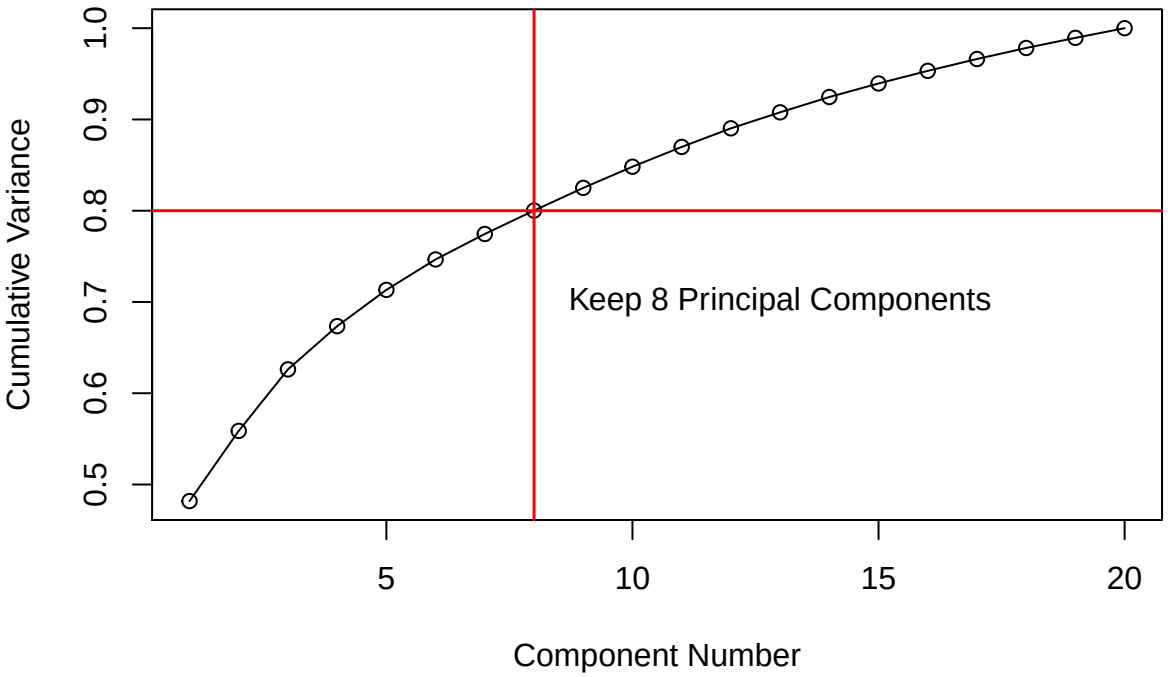

Figure 6: Scree cumulative variance explained

3.2. Predictive model with Principle Components as predictors

The scores of the principle components are used as predictors. Score is calculated by multiplying the variable values minus the variable's value by the loading of the variable.

We will split the scores into training and test sets for modelling. 70% of the data is chosen in random as the training set and the remaining of data is used as validation set. Ideally, the principle components' eigenvectors are computed based on the training sample and is used to compute the principle components of the validation or test sample. For simplicity, in this assignment the priciple components are calculated using the entire data set.

Table 2: Training and Validation sampling

Data	Samples
Training set	367
Validation set	134
Total	501

3.2.1 Linear Regression with principle components as predictors

A linear regression model with the 8 principle component scores as perdictors is fit. The model and the fit is summarized below. Let us call the model, "PCA model".

VV = 0.00075 - 0.00227 * Comp.1 + 0.00046 * Comp.2 + 5e-04 * Comp.3 + 0.00016 * Comp.4 - 0.00018 * Comp.5 - 8e-05 * Comp.6 + 4e-05 * Comp.7 - 0.00035 * Comp.8

```
##
## Call:
## lm(formula = VV ~ Comp.1 + Comp.2 + Comp.3 + Comp.4 + Comp.5 +
##     Comp.6 + Comp.7 + Comp.8, data = returns.scores)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0132078 -0.0016047  0.0000119  0.0016485  0.0147418
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.520e-04  1.222e-04   6.155 1.56e-09 ***
## Comp.1      -2.265e-03  3.936e-05 -57.548  < 2e-16 ***
## Comp.2       4.582e-04  9.850e-05   4.651 4.25e-06 ***
## Comp.3       5.032e-04  1.053e-04   4.781 2.31e-06 ***
## Comp.4       1.551e-04  1.255e-04   1.236   0.2171
## Comp.5      -1.816e-04  1.370e-04  -1.325   0.1856
## Comp.6      -8.223e-05  1.497e-04  -0.549   0.5830
## Comp.7       3.958e-05  1.635e-04   0.242   0.8088
## Comp.8      -3.544e-04  1.706e-04  -2.077   0.0383 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002735 on 492 degrees of freedom
## Multiple R-squared:  0.8724, Adjusted R-squared:  0.8703
## F-statistic: 420.5 on 8 and 492 DF,  p-value: < 2.2e-16
```

Table 3: VIF of Principle components

PC	VIF
Comp.1	1
Comp.2	1
Comp.3	1
Comp.4	1
Comp.5	1
Comp.6	1
Comp.7	1
Comp.8	1

It is seen that VIF is 1. This is not surprising because the principle components are by design are orthogonal to each other. The eigenvectors are constrained to be orthogonal to each other.

3.2.2. Predictive metric (MAE) for PCA model

The table below shows the predictive performance of the PCA model. The training and test sample's mean absolute error (MAE) are close.

Table 4: MAE of PCA model

Model	Train.MAE	Test.MAE
PCA model	0.0020201	0.0019084

3.3. Linear regression models with tickers as predictors

It will be interesting to compare the performance of the "PCA model" with linear regression models using raw tickers as the predictors. We will compare the PCA models by fitting an arbitarty model and a full model like the models in section 2.2. To compare the arbitary and full models with PCA model, it would be useful to split the data set into training and validation samples.

Table 5: Training and Validation sampling of Raw returns

Data	Samples
Training set	134
Validation set	367
Total	501

3.3.1. Arbitary model

The results for the arbitary model is shown below.

VV = 4e-05 + 0.11552 * GS - 0.00219 * DD + 0.03199 * DOW + 0.20029 * HON + 0.02764 * HUN + 0.02047 * JPM + 0.1659 * KO + 0.1465 * MMM + 0.13541 * XOM

```
##
## Call:
## lm(formula = VV ~ GS + DD + DOW + HON + HUN + JPM + KO + MMM +
```

```
##     XOM, data = returns.train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0064406 -0.0015686 -0.0001513  0.0016071  0.0072019
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.983e-05  2.238e-04   0.178  0.85903
## GS           1.155e-01  2.375e-02   4.865 3.41e-06 ***
## DD          -2.185e-03  3.425e-02  -0.064  0.94923
## DOW          3.199e-02  2.146e-02   1.491  0.13862
## HON          2.003e-01  3.579e-02   5.596 1.34e-07 ***
## HUN          2.764e-02  1.448e-02   1.909  0.05857 .
## JPM          2.047e-02  1.986e-02   1.030  0.30487
## KO           1.659e-01  2.799e-02   5.928 2.84e-08 ***
## MMM          1.465e-01  5.231e-02   2.800  0.00592 **
## XOM          1.354e-01  3.324e-02   4.074 8.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002534 on 124 degrees of freedom
## Multiple R-squared:  0.891,  Adjusted R-squared:  0.8831
## F-statistic: 112.7 on 9 and 124 DF,  p-value: < 2.2e-16
```

Table 6: MAE of Arbitary model

Model	Train.MAE	Test.MAE
Arbitary model	0.0019327	0.0023605

3.3.2 Full model

A full model with all tickers as predictors is fit. The model and the results are shown below.

VV = -0.00491 + 0.01878 * AA + 0.01674 * BAC - 0.03598 * BHI + 0.05553 * CVX - 0.03835 * DD + 0.02859 * DOW + 0.05485 * DPS + 0.06792 * GS + 0.01405 * HAL + 0.02895 * HES + 0.16527 * HON + 0.01377 * HUN + 0.01242 * JPM + 0.11238 * KO + 0.11243 * MMM + 0.0127 * MPC + 0.01837 * PEP + 0.05303 * SLB + 0.06775 * WFC + 0.057 * XOM + 0.00585 * u

```
##
## Call:
## lm(formula = VV ~ ., data = returns.train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0067682 -0.0014185 -0.0001234  0.0015844  0.0061887
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004905   0.002054  -2.389 0.018583 *
## AA           0.018779   0.019075   0.984 0.327000
## BAC          0.016741   0.018488   0.906 0.367125
```

```
## BHI         -0.035978   0.017903  -2.010 0.046873 *
## CVX          0.055533   0.039762   1.397 0.165282
## DD          -0.038353   0.032560  -1.178 0.241327
## DOW          0.028591   0.021594   1.324 0.188195
## DPS          0.054851   0.028097   1.952 0.053415 .
## GS           0.067924   0.024510   2.771 0.006541 **
## HAL          0.014055   0.021766   0.646 0.519779
## HES          0.028954   0.019835   1.460 0.147158
## HON          0.165270   0.034700   4.763 5.76e-06 ***
## HUN          0.013772   0.014262   0.966 0.336296
## JPM          0.012417   0.019005   0.653 0.514856
## KO           0.112375   0.033173   3.388 0.000974 ***
## MMM          0.112433   0.049947   2.251 0.026334 *
## MPC          0.012701   0.012154   1.045 0.298255
## PEP          0.018365   0.040165   0.457 0.648380
## SLB          0.053032   0.028543   1.858 0.065801 .
## WFC          0.067751   0.030623   2.212 0.028965 *
## XOM          0.057005   0.036159   1.576 0.117735
## u            0.005849   0.002418   2.419 0.017155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0023 on 112 degrees of freedom
## Multiple R-squared:  0.9189, Adjusted R-squared:  0.9037
## F-statistic: 60.44 on 21 and 112 DF,  p-value: < 2.2e-16
```

Table 7: MAE of Full model

Model	Train.MAE	Test.MAE
Full model	0.0015884	0.0034307

3.3.3 Model comparison

THe models that have been fit so far is shown in the table below.

```
##
## ---------------------------------------
## Models   Equation
## -------- ------------------------------
## PCA      VV = 0.00075 - 0.00227 *
##          Comp.1 + 0.00046 * Comp.2 +
##          5e-04 * Comp.3 + 0.00016 *
##          Comp.4 - 0.00018 * Comp.5 -
##          8e-05 * Comp.6 + 4e-05 *
##          Comp.7 - 0.00035 * Comp.8
##
## Arbitary VV = 4e-05 + 0.11552 * GS -
##          0.00219 * DD + 0.03199 * DOW +
##          0.20029 * HON + 0.02764 * HUN
##          + 0.02047 * JPM + 0.1659 * KO
##          + 0.1465 * MMM + 0.13541 * XOM
##
## Full     VV = -0.00491 + 0.01878 * AA +
```

```
##          0.01674 * BAC - 0.03598 * BHI
##          + 0.05553 * CVX - 0.03835 * DD
##          + 0.02859 * DOW + 0.05485 *
##          DPS + 0.06792 * GS + 0.01405 *
##          HAL + 0.02895 * HES + 0.16527
##          * HON + 0.01377 * HUN +
##          0.01242 * JPM + 0.11238 * KO +
##          0.11243 * MMM + 0.0127 * MPC +
##          0.01837 * PEP + 0.05303 * SLB
##          + 0.06775 * WFC + 0.057 * XOM
##          + 0.00585 * u
## ---------------------------------------
##
## Table: Model equations
```

The predicitve perfromance of the models are compared in table 8. It is seen that from a predictive performance stand point the PCA model is the best model; it has the lowest MAE. Also, from a VIF perspective the PCA model has the lowest VIF possible, which makes the regression coefficients stable. This reflects in the predictive performance, especially when there test sample has minor extrapolation in the predicitve space compared to the training sample.

It is seen that not all principle components' coefficiennts where statistically signifiant, it'll be useful to see how automated variable selection methods choose principle components as predictors. PCA being an unsupervised variable reduction technique, it'll be useful to wrap this technique with supervised (variable selection with response variable) automated variable selection technique.

Table 8: Predictive performance of regression models

Model	Train.MAE	Test.MAE
PCA model	0.0020201	0.0019084
Arbitary model	0.0019327	0.0023605
Full model	0.0015884	0.0034307

4. Automated variable selection

```
##
## Call:
## lm(formula = VV ~ ., data = train.scores)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0133126 -0.0016534  0.0000235  0.0014727  0.0136223
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.453e-04  3.037e-04   1.795 0.073459 .
## Comp.1      -2.262e-03  4.642e-05 -48.739  < 2e-16 ***
## Comp.2       4.048e-04  1.160e-04   3.489 0.000548 ***
## Comp.3       4.738e-04  1.264e-04   3.748 0.000209 ***
## Comp.4       2.291e-04  1.437e-04   1.595 0.111712
## Comp.5      -1.717e-04  1.607e-04  -1.069 0.286007
## Comp.6      -1.841e-04  1.770e-04  -1.040 0.299031
## Comp.7       8.542e-05  1.879e-04   0.455 0.649743
```

```
## Comp.8      -2.705e-04  1.995e-04  -1.356 0.176088
## Comp.9      -4.584e-04  2.034e-04  -2.254 0.024820 *
## Comp.10      3.470e-04  2.097e-04   1.655 0.098884 .
## Comp.11      7.504e-04  2.164e-04   3.468 0.000591 ***
## Comp.12     -1.535e-05  2.206e-04  -0.070 0.944573
## Comp.13     -2.862e-04  2.462e-04  -1.162 0.245850
## Comp.14      4.377e-04  2.475e-04   1.768 0.077879 .
## Comp.15      9.629e-05  2.630e-04   0.366 0.714533
## Comp.16     -4.646e-04  2.804e-04  -1.657 0.098497 .
## Comp.17     -2.444e-04  2.938e-04  -0.832 0.406223
## Comp.18     -1.654e-04  2.885e-04  -0.573 0.566796
## Comp.19      2.469e-04  3.171e-04   0.779 0.436753
## Comp.20      1.499e-04  3.205e-04   0.468 0.640300
## u            5.697e-04  7.177e-04   0.794 0.427913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002769 on 345 degrees of freedom
## Multiple R-squared:  0.8771, Adjusted R-squared:  0.8696
## F-statistic: 117.3 on 21 and 345 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = VV ~ Comp.1 + Comp.2 + Comp.3 + Comp.4 + Comp.8 +
##     Comp.9 + Comp.10 + Comp.11 + Comp.14 + Comp.16, data = train.scores)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.013595 -0.001649  0.000062  0.001465  0.014160
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.521e-04  1.437e-04   5.232 2.87e-07 ***
## Comp.1      -2.261e-03  4.598e-05 -49.182  < 2e-16 ***
## Comp.2       4.058e-04  1.150e-04   3.530 0.000470 ***
## Comp.3       4.941e-04  1.245e-04   3.970 8.71e-05 ***
## Comp.4       2.163e-04  1.423e-04   1.520 0.129485
## Comp.8      -2.792e-04  1.971e-04  -1.417 0.157442
## Comp.9      -4.515e-04  2.009e-04  -2.247 0.025246 *
## Comp.10      3.242e-04  2.067e-04   1.568 0.117664
## Comp.11      7.228e-04  2.134e-04   3.387 0.000784 ***
## Comp.14      4.591e-04  2.428e-04   1.890 0.059507 .
## Comp.16     -4.311e-04  2.741e-04  -1.573 0.116554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002751 on 356 degrees of freedom
## Multiple R-squared:  0.8749, Adjusted R-squared:  0.8714
## F-statistic: 248.9 on 10 and 356 DF,  p-value: < 2.2e-16
```

```
##   Comp.1   Comp.2   Comp.3   Comp.4   Comp.8   Comp.9  Comp.10  Comp.11
## 1.004340 1.009935 1.009405 1.008870 1.004940 1.008710 1.005426 1.007111
##  Comp.14  Comp.16
## 1.006297 1.005230
```