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.
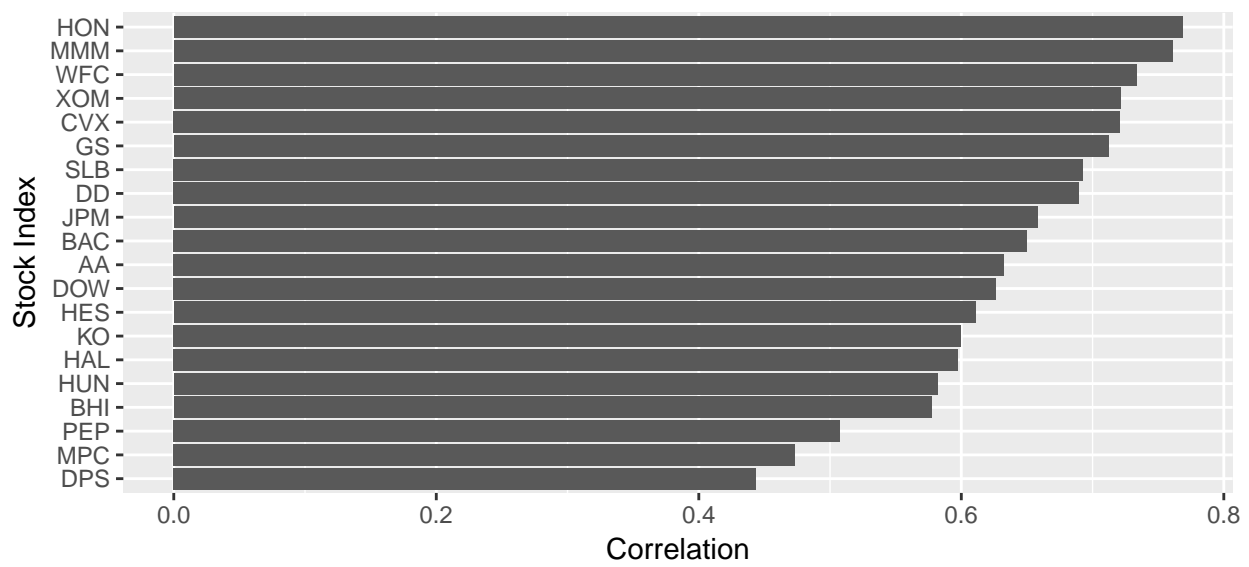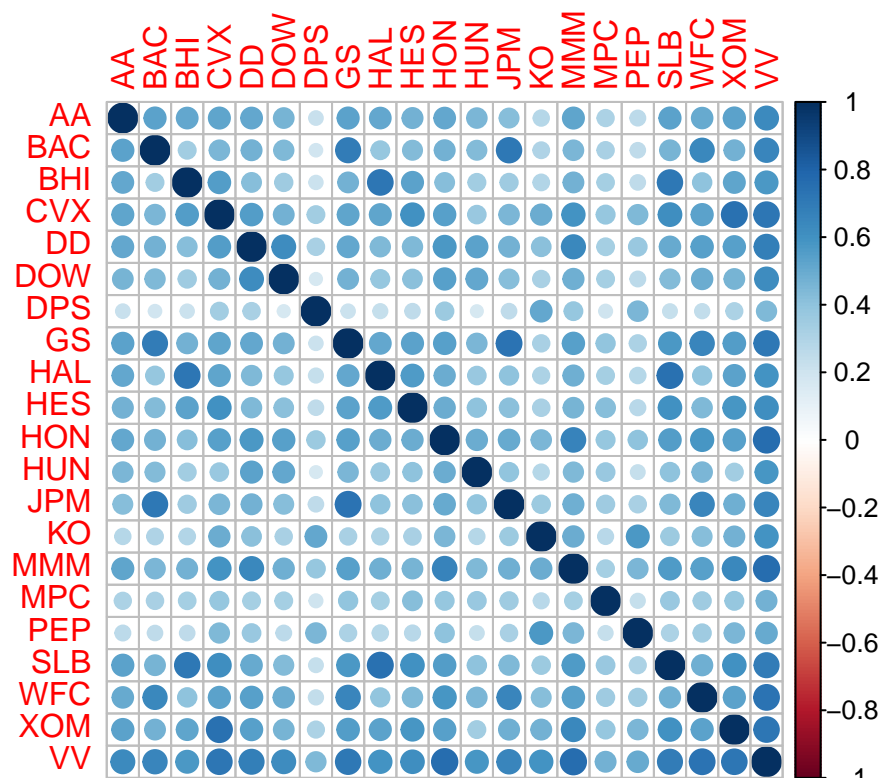
Figure 1: Correlations with VV

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. 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: VIF Full model

Model	Predictors	VIF
	GS	2.705795
	MMM	2.590177
Arbitary model	DD	2.368257
	JPM	2.324600
	HON	2.261397
	SLB	3.258982
	GS	3.197811
Full Model	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
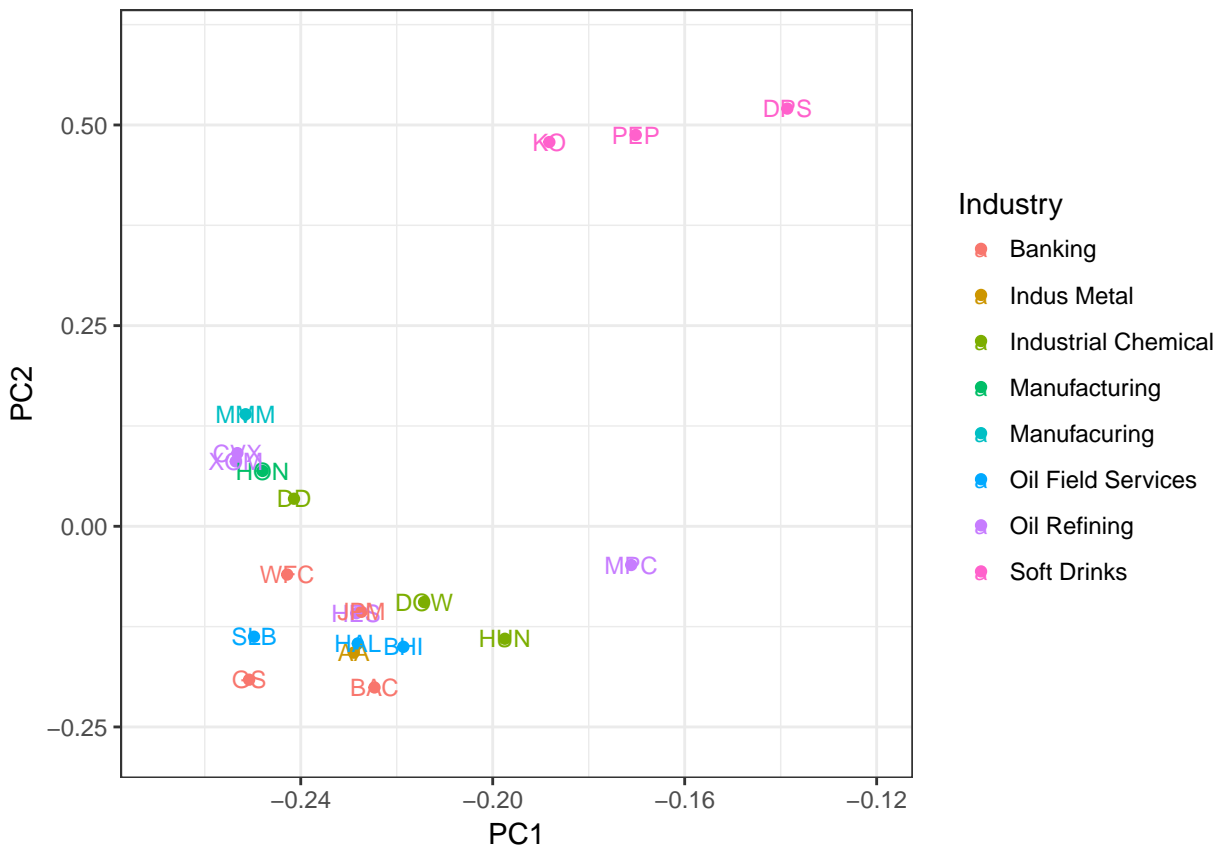
Figure 3: Variable loadings of first 2 principle components

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.

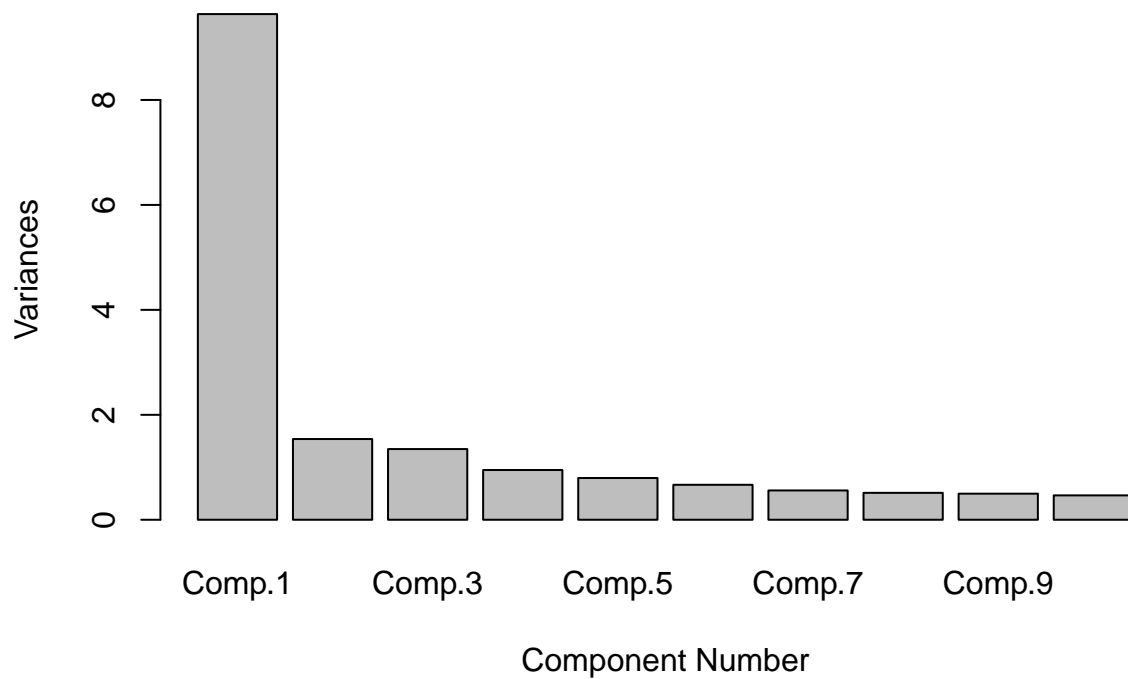Figure 6 shows the first 8 components explain 80% of the total variance in the data.

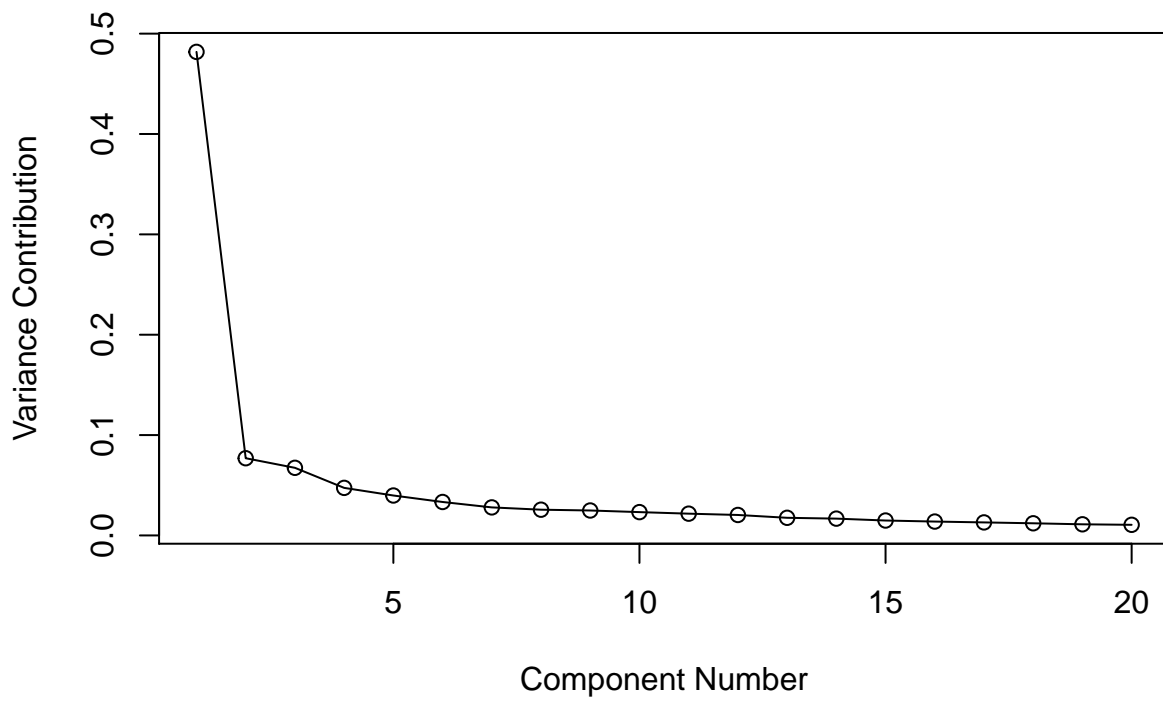
Figure 4: Scree Plot - Not so good looking
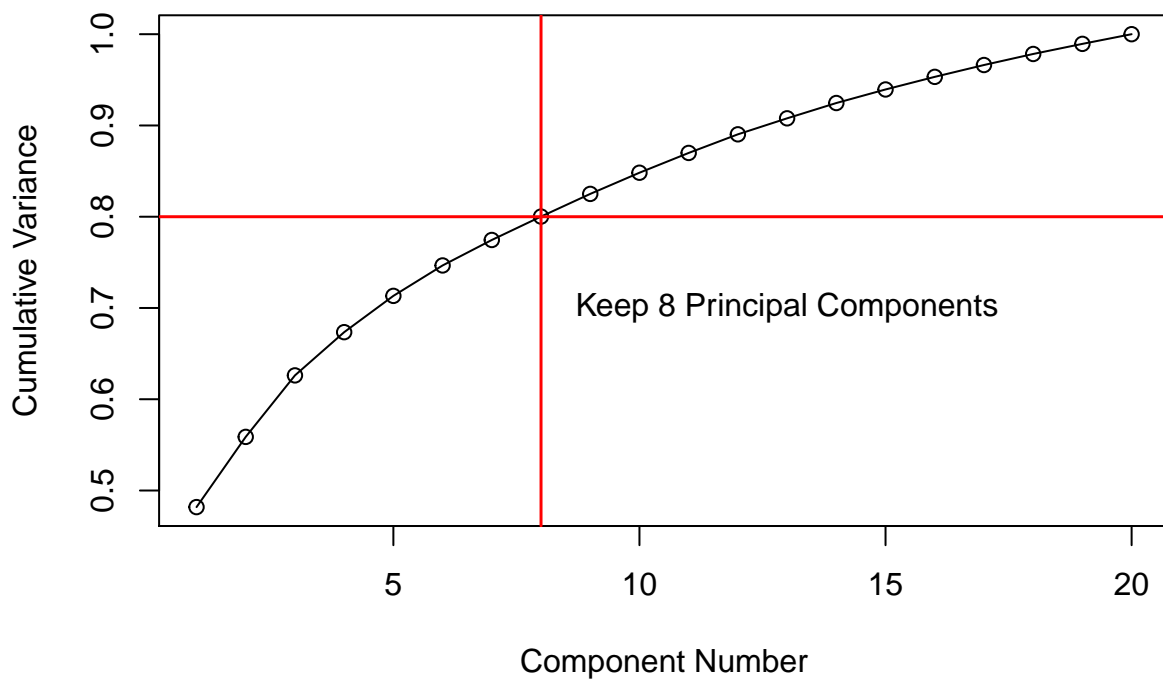
Figure 5: Scree plot - Variance explained

Keep 8 Principal Components

Figure 6: Scree cumulative variance explained