Assignment #6

Andrew Knight

Introduction:

The goal of this assignment is to perform multivariate analysis using applied examples while demonstrating knowledge of the concepts.

Data File: stock_portfolio.csv

Calculations File: Knight_Assignment6.R

Assignment Tasks

1: Data Preparation

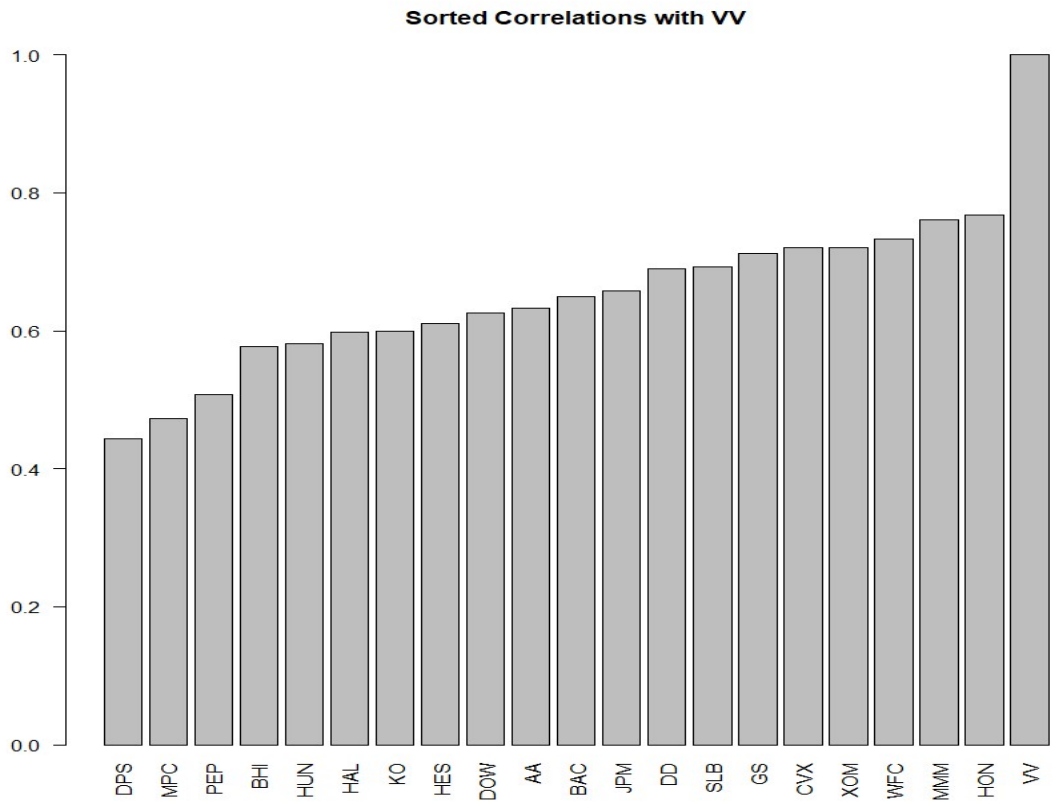
The stock portfolio data is loaded and basic review is performed. We have twenty stock symbols represented with daily closing prices for each as well as the price for a single index fund, VV as the response. A single row in the data frame contains the date (working weekdays only), the price for each stock and the price of the index fund. All variables with the exception of the date field contain continuous numeric data with two years of stock data (501 rows) starting with Jan 3rd, 2012.

2: EDA

First off, we can sort the dates and perform a transformation using the log of today's price versus yesterday's price. By doing this for each stock we take remove the autocorrelation that is inherent in the time series data. Time series data of incremental dates like we have in this dataset suffers from the highly correlated prices for subsequent days. Also, because this was performed for each stock referenced in the data, we need to do the same for the index fund to keep it consistent.

For our correlation plot, we first want to compare correlations for each of the twenty stocks with the index VV. After evaluating the cor() results and sorting them in decreasing order, we see that HON and MMM exhibit the highest correlation with VV and DPS and MPC exhibit the lowest.

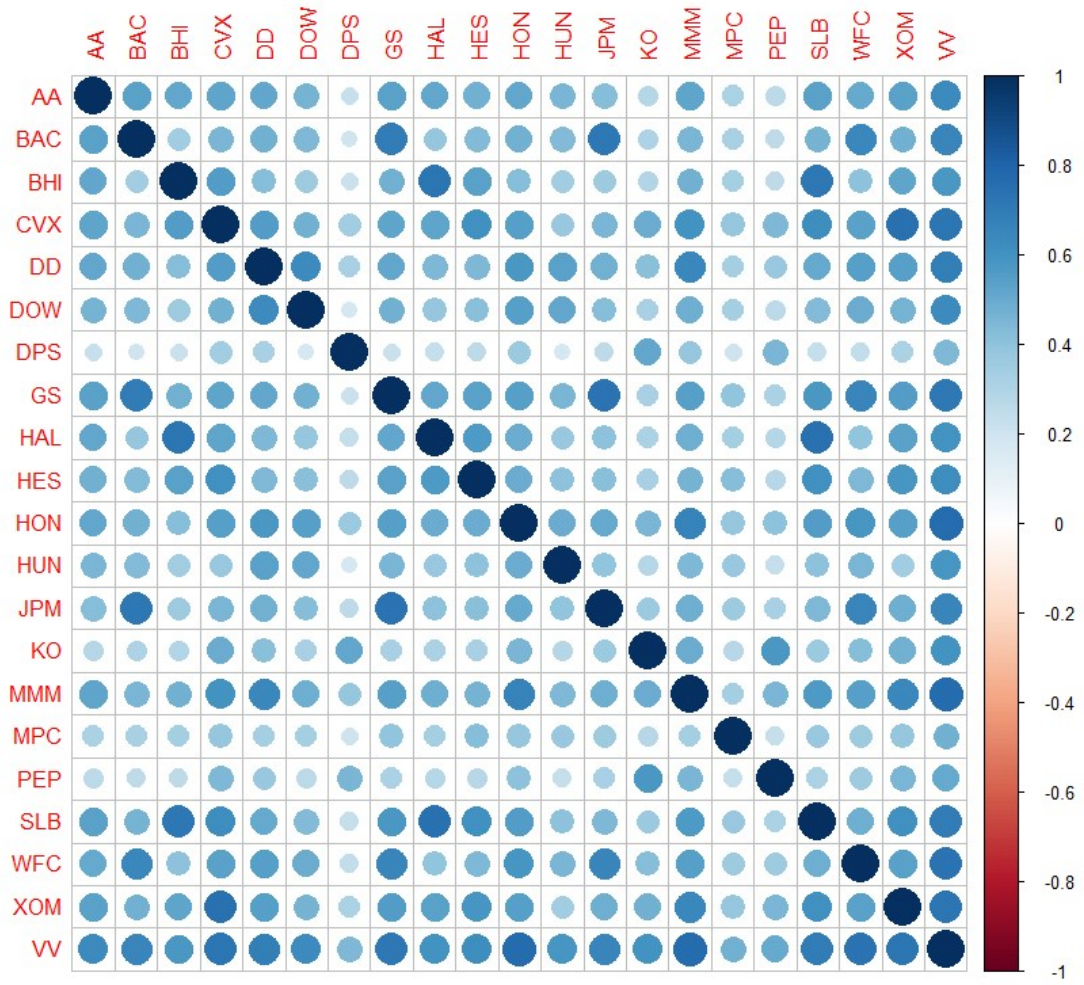
Another great way to compare each stock correlation factor is to view the sorted results with a barchart as shown below.



There are some additional things to note in this plot. First, the variation in the correlation factors do not vary all that much. Looking at the summary for these values we get a min of 0.4435 and a max of 0.7684. Also, the median value is 0.6502 for the stock BAC.

3: Visualization

Using corrplot, we can get a better visual representation of the correlation for each pair of stocks. The plot is shown here.



When analyzing this one of the first things I see is that we have no variables with negative correlation. This is expected because we are viewing the log-transformed returns values. If we did see some negatively correlated values, that would indicate something is wrong with our calculations.

I also note that the stocks with the lowest pairwise correlations to the other stocks show up as lighter bands for almost all other stocks. These include DPS, MPC, and PEP. This means that these three are not closely correlated with most other stocks. Looking specifically at the circles in the VV column, we will find the stocks that have the highest correlation with the index fund. Based on the visual clues alone, going with CVX, DD, GS, HON, MMM, SLB, WFC, and XOM may be a good selection of variables for the model. We will review possible models in the next section.

Before looking at the VIF scores in R, we would expect DPS, MPC and PEP to have the lowest VIF values. Similarly, based on the plot above I would expect HON, MMM, and WFC to have the highest values. These models will be compared further in the next section.

4: Models

Now we need to look at a couple possible models and review the actual VIF scores. For this I tried three models to start. The first model was a limited model using only HON, MMM, and WPC. The second model was the 'middle' test to check the 'good' list of stocks based on visual analysis of correlated variables above. The final model uses all available variables.

Here is the comparison between the three models.

Model	RSE	R-squared	F-stat	Highest VIF
Limited	0.00361	0.7755	572.2	2.068
Middle	0.00300	0.8460	337.9	2.766
Full Model	0.00267	0.8818	179.0	3.259

From this table data, we notice that characteristics (such as RSE and R-squared) of the model improves as we add more variables. However, we also see that the VIF scores increase. I've only listed the highest VIF values however the values across the board for the min, median and max scores overall increase as we add more variables to the model moving from limited up to the full model. This is a clear indication of the existence of collinearity. The F-statistic continues to decrease as variables are added as well.

If we were to continue this trend to build a larger model of more stock predictors we could quickly find ourselves in scenario where we begin to have serious issues determining the correct variables to include in our model to predict the index fund response. If we were to try to add an additional 20 stocks to the mix it will be very difficult to find the right model without a means of dimension reduction.

5: Principal Component Analysis

The purpose of this section is to look at the principal components to see how we can use them as the predictors for a model that does not suffer from multicollinearity.

```
> summary(returns.pca)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11
Standard deviation     3.1042633 1.24037473 1.16079159 0.97348817 0.89191173 0.8163381 0.74727540 0.71606462 0.70486968 0.68141987 0.65836107
Proportion of Variance 0.4818225 0.07692647 0.06737186 0.04738396 0.03977533 0.0333204 0.02792103 0.02563743 0.02484206 0.02321665 0.02167196
Cumulative Proportion  0.4818225 0.55874901 0.62612087 0.67350483 0.71328015 0.7466005 0.77452158 0.80015900 0.82500107 0.84821772 0.86988968
                         Comp.12   Comp.13    Comp.14    Comp.15    Comp.16    Comp.17   Comp.18   Comp.19    Comp.20
Standard deviation     0.6385577 0.5925250 0.57888423 0.54494939 0.52563057 0.50927436 0.4919940 0.4710181 0.46013436
Proportion of Variance 0.0203878 0.0175543 0.01675535 0.01484849 0.01381437 0.01296802 0.0121029 0.0110929 0.01058618
Cumulative Proportion  0.8902775 0.9078318 0.92458713 0.93943562 0.95324999 0.96621801 0.9783209 0.9894138 1.00000000
```

The princomp results give us the principal components for each of our 20 variables excluding the VV response. Because principal components are ordered, the first components are what we are most interested in. The summary from this function gives us the standard deviation, proportion of variance and cumulative proportion for each component. The variable loadings is the matrix of the eigenvectors for each component and from the first column we can form the first principal component for each stock. When we plot the first component versus the second component, we can see these eigenvector values and see which of the stocks have a higher correlation between the respective components. There are a

few trends when comparing each of the companies. The first two components show higher correlation for soft drink beverage companies whereas the banking and oil field services companies show a weaker correlation between them.

6: Component Selection

What we are really interested in, however, is the scree plot which gives us a clear indication of which principal components provide the most variance in the component data. In our case pc1 explains almost 50% of the variation before a big drop down to pc2 at less than 8%. Over 70 percent of the total variation is covered by the first five principal components. If we want to determine a cutoff percentage, the total variance plot can show the cumulative variance for each successive component.

The number of principal components you decide to keep is largely a decision that must be made with goals of the particular project in mind. The ultimate goal is of course to determine a model which performs well for out-of-sample data, but we also aim to reduce the amount of collinearity enough so that we can accurately explain the independent predictors and their contribution to the response. To do this, we need an 'acceptable' level of non-collinearity. Based on our analysis thus far, it seems the middle of the road model from above consisting of eight principal components is a good first choice. We can see that they explain 80% of the total variance. Combined with the somewhat arbitrary max VIF cutoff of 3.0, it seems to be the best choice among the three listed so far.

7: Predictive Scores

The principal component scores contain the mapping of the original data into their principal components for all of the twenty variables. The score for each is essentially a numerical value indicating the strength the contribution for that component.

The data was split into train and test samples using random assignment with 70% in the training set. The RSE and R^2 results of the linear model are very comparable to the models using original data above in section 4. Using the first eight principal components, we've computed the linear model with a RSE of 0.00275 and an R-squared value of 0.863 however looking at the p-values we see that only the first four are principal components are statistically significant. The VIF values are all approximately 1 as we expected from the PCA which now means we have no collinearity. By definition, all principal components are orthogonal and thus give the lowest possible VIF value of one.

We also calculate the mean absolute error for the pca1 model with the first eight components and we get 0.00200 and 0.00198 for the train and test sets respectively. The in-sample MAE (training set) is 0.00199 and the out-of-sample MAE (test set) is 0.00195 which seems reasonable.

8: Model Comparison

Now we run our three models again using the original data separated into training and test samples and we get the following results. Note that the RSE and R^2 values are the same as listed in table above.

Model	RSE	R-squared	MAE Train	MAE Test
Limited	0.00361	0.7755	0.00761	0.00270
Middle	0.00300	0.8460	0.00770	0.00222
Full Model	0.00267	0.8818	0.00783	0.00184

One important thing to note here is that the RSE, R-squared, and MAE values are all very similar across the three models. Especially when comparing the middle and full models, we see that similar predictive performance can be achieved with fewer variables using PCA.

Ultimately, the best model definition should come down to creating a model that hits the desired goals with the fewest number of variables. This approach helps us streamline our model selection process through model specs, parameter estimation, adequacy checking and validation.

9: Best Model

The PCA model should not necessarily be selected as the best model however it is a useful transformation to have for addressing the multicollinearity issue. The unsupervised learning aspect of PCA allows us to choose a number of predictors based on the amount of variation explained. In this way I view this characteristic as being unsupervised because it does not rely on the number of predictors to make a decision. We are simply interested in cumulative percentage of variance. By choosing an overall percentage and then testing the model with the number of principal components required to achieve that percentage, we can verify the performance of the components directly without considering the original variables. When it comes time to compare the predictive performance of the component versus the original variables, we can then apply the 'supervised' aspect of the analysis. By selecting the PCA model that has similar fit performance as the original, we can verify that the PCA model achieves the intended goal of matching error and R^2 values but with less variables. Using the opposite approach of starting with the analysis of which original variables to keep would be difficult with collinearity present and the pairwise analysis of variables in a visual way could be prohibitive as the number of predictors grows. When attempting a forward stepAIC on various untransformed models, I was having difficulty narrowing in on the right reduced model without using unsupervised PCA selection.

Our goal for choosing the 'best' model should be to obtain a reduced-variable model that achieves similar predictive performance as a full model using original data. Based on our models above, the decision to include eight is based on the fact that we can account for 80% of the variation by choosing the first eight principal components. We can achieve similar predictive performance (an R^2 of about

0.87) fitting the top principal components while reducing the complexity of the model. This variable reduction technique not only resolves the multicollinearity problem, it gives us confidence in our analysis that we are not missing important data in higher dimensional problems. This is important because as the size of the variables grows, our ability to properly perform pairwise comparisons using traditional univariate methods becomes costly, time consuming or outright impossible to complete in a structured manner.

Using the backward stepwise AIC process on each model, we obtain the following MAE values.

Model	**R-squared**	**MAE Train**	**MAE Test**
PCA Model – back AIC	0.8632	0.00198	0.00195
Orig Model – back AIC	0.8699	0.00783	0.00184

The PCA model uses variables Comp1, Comp2, Comp3, Comp4, Comp8, Comp9, Comp10, Comp11 and has an overall R-squared values 0.8632. The Original Model using the full list of predictors has an overall R-squared value of 0.8699. Thus, the predictive performance is similar but with fewer variables. I saw comparable results in the RSE, R-squared, and MAE values are all very similar across the various models tested. Especially when comparing the middle and full models, we see that similar predictive performance can be achieved with but with fewer variables using PCA.

One interesting observation was the increase in the MAE value for the test set for the PCA model over the original model. This may be attributed to the randomized split used for the train and test set but in our case, it seems like an acceptable tradeoff for the results obtained. Of course, the benefit of using the PCA model is that the VIF values are all very close to 1 whereas the VIF scores for the original models averaged between 2 and 3 depending on the predictors. This may not be a huge issue in our case but as the number of stock listings increase that we want to compare, this could become a problem.

It seems clear that the PCA model provide the most benefits and should be selected as the 'best model' to use. Reducing the variables and removing the multicollinearity without compromising the R-squared were the main reasons for choosing it. The model using principal components has a lower mean average error and allow us to focus on the individual contributions for each predictor to find the find the most meaningful in fitting the regression model.

Reflections & Conclusion

In summary, principal component analysis gives us another tool for the EDA and model selection process when working with multivariate data. It can be used as a variable reduction method to address multicollinearity. However, any resulting 'best' model must still be validated and 'scored' against the original modeling goals of the organization. Proper data cleaning, diagnostics and outlier analysis is still required. It is aimed at making the variable selection process part of model specification easier and less

reliant on visual analysis for each variable. Tools like R make it easy to try different models including PCA and to quickly hone in on the desired outcome.