Noé Flores
Factor Analysis

Introduction:

This analysis will aim to utilize factor analysis to identify sectors in the stock. The initial data set is made up of daily closing prices for 20 stocks and a large-cap index ETF from Vanguard. For the purpose of this analysis, we will drop some of these variables in order to achieve better factor analysis results.

Results and Analysis:

Our examination begins by making the previously described adjustments to our data. We will be removing the following symbols from the data set: AA, HON, MMM, DPS, KO, PEP, MPC, and GS. This will leave us with specific market sectors consisting of (1) Banking, (2) Oil Field Services, (3) Oil Refining, and (4) Industrial – Chemical. We can then hypothesize that we will then have three to four factors (industry sectors) in this data set. The core of our examination centers on the use of daily returns for the remaining stocks in the data set to arrive at our observations. Specifically, we will examine the log-returns of the individual stocks. The table below provides a brief example of the data and the transformed returns for four of the stocks and our index.

Figure 1: Closing Prices and Returns Example.

Closing Prices					
Date	**BAC**	**BHI**	**CVX**	**DD**	**VV**
3-Jan-12	5.8	51.02	110.37	46.51	58.18
4-Jan-12	5.81	51.53	110.18	47.02	58.25
5-Jan-12	6.31	50.82	109.1	46.70	58.44
Returns					
Date	**return_BAC**	**return_BHI**	**return_CVX**	**return_DD**	**response_VV**
3-Jan-12	.	.	.	.	.
4-Jan-12	0.001723	0.009946	-0.00172297	0.010906	0.00120244
5-Jan-12	0.082555	-0.013874	-0.0098505	-0.006829	0.00325649

With the data prepared, we can continue with our factor analysis. Our initial factor analysis will be performed through a SAS procedure that will automatically select the number of factors to retain for us. Factor analysis tries to explain the covariance or correlations of the observed variables by attempting to choose a small amount of latent variables or common factors, to which the more obvious variables are related. Factor analysis operates on the reduced covariance matrix. To calculate this reduced covariance matrix, we need values for estimated commonalities that are calculated from the estimated factor loadings. Given the initial commonalities, a principal component analysis is performed on the reduced covariance matrix, and the first eigenvectors are used to provide the estimate of the loadings in the factor model. Below is a breakdown of the Eigenvalues of the reduced covariance matrix.

Figure 2: Eigenvalues of Reduced Covariance matrix.

	Eigenvalues of the Reduced Covariance Matrix			
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	6.04732583	5.16261770	0.8812	0.8812
2	0.88470813	0.52262870	0.1289	1.0101
3	0.36207942	0.05735386	0.0528	1.0629
4	0.30472556	0.29429115	0.0444	1.1073
5	0.01043441	0.06365245	0.0015	1.1088
6	-0.05321803	0.01517115	-0.0078	1.1011
7	-0.06838918	0.03291807	-0.0100	1.0911
8	-0.10130725	0.01600696	-0.0148	1.0763
9	-0.11731422	0.00866270	-0.0171	1.0593
10	-0.12597692	0.01040221	-0.0184	1.0409
11	-0.13637913	0.00786652	-0.0199	1.0210
12	-0.14424565		-0.0210	1.0000
Total = 6.86244298		**Average = 0.57187025**		

The table in Figure 2 shows that the first two eigenvalues account for a significant amount of the variance, likely due to the fact that variables in our model are all highly correlated and there is some latent variable to which the more obvious variables are related. From this table, we can also determine how many factors will be chosen and why. Two factors will be retained, based on the default option in SAS which follows specific criteria if no prior minimum is specified. The MINEIGEN rule specifies that the smallest eigenvalue for which a factor is retained will be based on the following calculation:

$$\text{MINEIGEN} = \frac{Total\ Weighted\ Variance}{Number\ \ of\ Variables}$$

In our case, the results of the equation are given at the bottom of the table and therefore we can see that two eigenvalues are higher than the average eigenvalue of .57, which leads to the retention of two factors.

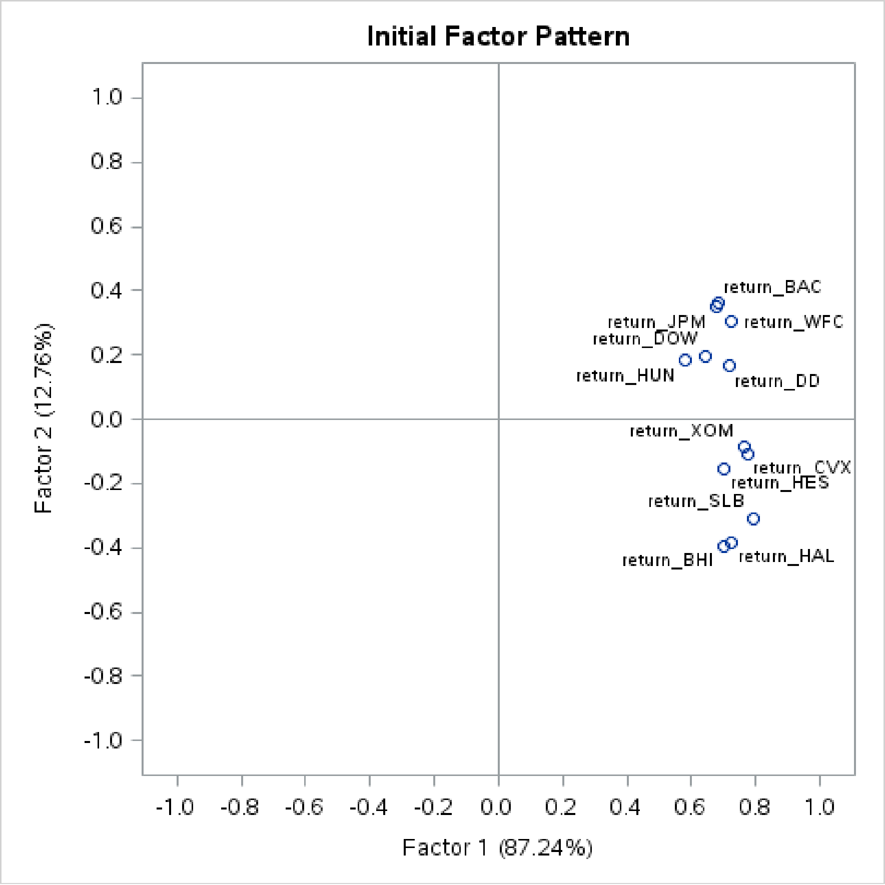
Figure 3: Factor Patter.

Factor Pattern		
	Factor1	**Factor2**
return_BAC	0.68475	0.36021
return_BHI	0.69984	-0.39498
return_CVX	0.77402	-0.10833
return_DD	0.71605	0.16703
return_DOW	0.64548	0.19801
return_HAL	0.72630	-0.38221
return_HES	0.70361	-0.15709
return_HUN	0.58030	0.18186
return_JPM	0.67874	0.34813
return_SLB	0.79382	-0.30815
return_WFC	0.72445	0.30517
return_XOM	0.76500	-0.08361
Variance Explained by Each Factor		
	Factor1	**Factor 2**
	6.0473258	0.8847081
Final Communality Estimates:		
		Total = 6.932034

In Figure.3 we see the initial un-rotated factor structure matrix, consisting of the correlations between the 12 subtests and the two factors retained by SAS. The two factors that were retained by SAS fall short and fail to support the initial hypothesis of three to four factors. Our results don't appear to show simple structure, meaning that most items have a large loading on one factor but smaller loadings on other factors. It is a bit strange to see that the variables are correlated so highly for the first factor and not as much for the second factor, exhibiting negative correlation values.

If we examine the two factors graphically, we can see that there appears to be a bit of clustering in our variables. An interesting note on the graph is that one cluster is composed of BAC, JPM, WFC, DOW, HUN, and DD, while the second cluster is composed of XOM, CVX, HES, SLB, BHI, and HAL. These clusters align with the values we observed in Figure.3 and appears to be grouping the variables by Banking and Industrial sectors in one cluster vs. Oil Field services and Oil Refining sectors in the second. Overall, it is slightly difficult to interpret the results of the non-rotated factor analysis.

Figure 4: Factor.1 and Factor.2 Plot.



We continue by executing a factor analysis with rotation. Factor rotation will allow us to clarify and simplify the factor analysis as much as possible. It does not alter the overall structure of the solution. This makes the answer more interpretable without changing the underlying mathematical properties. The use of rotation should make the output more distinct and give us the simple factor structure we were missing in the first analysis. We will use Varimax rotation, with the aim of this method being for factors with few large loadings and as many near-zero loadings as possible. It is intended to produce factors that have correlations with one small set of variables and little to no relationship with another set.

Figure 5: Rotated Factor Patter.

Rottated Factor Pattern		
	Factor1	**Factor2**
return_BAC	0.73912	0.22875
return_BHI	0.21634	0.77394
return_CVX	0.47133	0.62344
return_DD	0.62482	0.38759
return_DOW	0.59675	0.31582
return_HAL	0.24408	0.78359
return_HES	0.38705	0.60822
return_HUN	0.53921	0.28120
return_JPM	0.72634	0.23305
return_SLB	0.34419	0.77886
return_WFC	0.72835	0.29575
return_XOM	0.48241	0.59958
Variance Explained by Each Factor		
	Factor1	**Factor 2**
	3.4711423	3.4608916
Final Communality Estimates:		
		Total = 6.932034

SAS once again chose two factors, but there is a noticeable difference in the rotated factor values. The components of the factors are the same as our first analysis, but the interpretability is substantially improved. Factor.1 and Factor.2 each appear to explain the same amount of variance making it easier to conceptualize. We also have a simple factor structure in place. Most of our variables now have a large loading on one factor but smaller loadings on the other factor. This also makes it easier to interpret which stocks make up each factor. The overall interpretability has been significantly improved.

Another method commonly used in the estimation of common factors is the maximum likelihood factor analysis (ML). The use of the maximum likelihood method requires us to assume that the data is independently sampled from a multivariate normal distribution and is intended to define a distance measure between the observed covariance matrix, and the covariance matrix implied by the factor analysis model.

Figure 6: Eigenvalues of the Weighted Reduced Covariance Matrix.

	Eigenvalues of the Weighted Reduced Covariance Matrix			
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	16.1720778	13.4481419	0.8558	0.8558
2	2.7239360	1.9046476	0.1442	1.0000
3	0.8192884	0.2494707	0.0434	1.0434
4	0.5698176	0.4658764	0.0302	1.0735
5	0.1039412	0.1141060	0.0055	1.0790
6	-0.0101647	0.0295222	-0.0005	1.0785
7	-0.0396869	0.0990329	-0.0021	1.0764
8	-0.1387198	0.1324675	-0.0073	1.0690
9	-0.2711873	0.0110888	-0.0144	1.0547
10	-0.2822761	0.0481722	-0.0149	1.0397
11	-0.3304483	0.0901170	-0.0175	1.0223
12	-0.4205653		-0.0223	1.0000
Total =	**18.8960127**	**Average = 1.57466772**		

Based on the initial results of the weighted reduced covariance matrix, SAS will once again suggest the retention of two factors. The MINEIGEN rule specifies that the smallest eigenvalue for which a factor is retained will be based on the following calculation:

$$\text{MINEIGEN} = \frac{Total\ \ Weig\ hted\ Variance}{Number\ \ of\ Variables}$$

In our case, the results of the equation are given at the bottom of the table and therefore we can see that two eigenvalues are greater than the average of 1.575. One of the benefits of utilizing a maximum likelihood method is the fact that it allows us to test the hypothesis about the number of common factors. The hypothesis test information in Figure.7 below corroborates with our results and allows us to accept the null hypothesis of two factors.

Figure 7: Hypothesis test.

Significance test based on 501 Observations			
Hypothesis Test	**DF**	**Chi-Square**	**Pr > Chi-Sq**
H0: No common factors	66	3656.2617	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	43	319.3192	<.0001
HA: More factors are needed			

Figure 8: Rotated Factor Pattern ML.

Rotated Factor Pattern		
	Factor1	**Factor2**
return_BAC	0.76122	0.21969
return_BHI	0.21664	0.79932
return_CVX	0.49806	0.57530
return_DD	0.59542	0.38748
return_DOW	0.56395	0.31884
return_HAL	0.24256	0.80907
return_HES	0.40289	0.59153
return_HUN	0.50588	0.29457
return_JPM	0.75054	0.22277
return_SLB	0.35223	0.79376
return_WFC	0.75994	0.27534
return_XOM	0.51113	0.55362

Variance Explained by Each Factor		
Factor	**Weighted**	**Unweighted**
Factor1	8.7156851	3.55022275
Factor2	10.1803287	3.42320994

There doesn't appear to be a significantly noticeable difference in the rotated factor values and those in the maximum likelihood estimation. The components of the factors remain the same as our first analysis, but the interpretability here is also substantially improved over a basic principal factor analysis. Factor 1 and Factor 2 don't quite explain the same amount of variance, but there are some differences in this modeling technique that provide a more robust perspective. The treatment of the correlations is weighted by their uniqueness, and we also have a number of fit indices which are displayed below.

Figure 9: Goodness of Fit.

Goodness of Fit	
Chi-Square without Bartlett's Correction	323.30664
Akaike's Information Criterion (AIC)	237.30664
Schwarz's Bayesian Criterion (SBC)	55.99257
Tucker and Lewis's Reliability Coefficient	0.88187

Chi-square, which assesses the goodness of fit between observed and theoretical values is displayed with and without Bartlett's Correction. The (AIC) is a general measure for estimating the best number of parameters to include in a model when maximum likelihood estimation is used. Typically, the number of factors that yields the smallest value of AIC is considered best. The SBC is another gauge for determining the best number of parameters. Once again, the number of factors that yields the smallest value of SBC is considered to be the best. The Tucker and Lewis’s Reliability Coefficient is an indicator of reliability, with value ranges from 0 to 1. A larger value indicates better reliability. These metrics are all useful for the model goodness of fit comparisons.

As a final Factor analysis procedure, we will look to change the estimates of the prior commonalities by adjusting our priors option to MAX. Making this adjustment should make a significant difference in the number of factors retained.

Figure 10: Eigenvalues of the Weighted Reduced Covariance Matrix.

Eigenvalues of the Reduced Covariance Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	21.0077686	17.5450008	0.7550	0.7550
2	3.4627678	1.7607252	0.1245	0.8795
3	1.7020426	0.3456833	0.0612	0.9406
4	1.3563592	1.0611028	0.0487	0.9894
5	0.2952564	0.1527889	0.0106	1.0000
6	0.1424675	0.0574075	0.0051	1.0051
7	0.0850600	0.0781812	0.0031	1.0082
8	0.0068788	0.0266398	0.0002	1.0084
9	-0.0197610	0.0152328	-0.0007	1.0077
10	-0.0349938	0.0478784	-0.0013	1.0065
11	-0.0828723	0.0139147	-0.0030	1.0035
12	-0.0967870		-0.0035	1.0000
Total =	**27.8241868**	**Average =**	**2.31868224**	

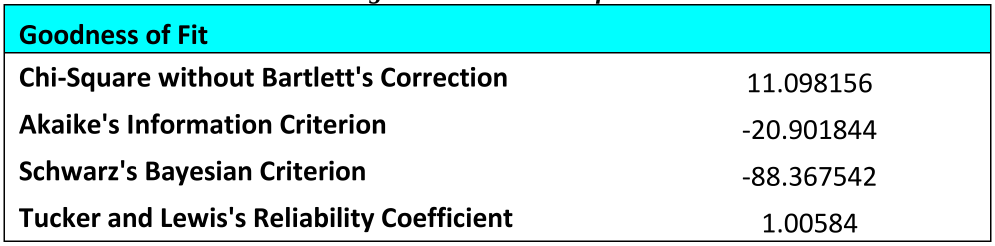
Based on the initial results of the weighted reduced covariance matrix, I would assume that SAS will once again suggest the retention of two factors, based on our formula and given the average of 2.319. However, the analysis has selected five total factors for retention.

Figure 11: Eigenvalues of the Weighted Reduced Covariance Matrix.

Rotated Factor Pattern					
	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
return_BAC	0.19300	0.75425	0.26803	0.17215	0.09285
return_BHI	0.75597	0.14970	0.18684	0.24628	-0.01722
return_CVX	0.37688	0.25354	0.2644	0.70383	0.02658
return_DD	0.24372	0.27524	0.66859	0.31138	-0.13337
return_DOW	0.19396	0.25931	0.64481	0.23505	-0.00701
return_HAL	0.82071	0.18978	0.20801	0.16916	-0.00609
return_HES	0.47834	0.23976	0.25785	0.40900	0.24903
return_HUN	0.22592	0.26677	0.60996	0.06709	0.16770
return_JPM	0.20547	0.77151	0.22874	0.17842	-0.03102
return_SLB	0.72537	0.25575	0.24707	0.30301	0.05701
return_WFC	0.20847	0.61032	0.35934	0.29285	-0.00631
return_XOM	0.37166	0.29603	0.24083	0.66560	-0.02404
Variance Explained by Each Factor					
Factor1	9.48177257	2.55119512			
Factor2	6.95572063	2.0840043			
Factor3	5.26449075	1.8217392			
Factor4	5.8023705	1.59069819			
Factor5	0.31984016	0.12246466			

The results suggest that the factor selection is greatly influenced by the sensitivity of the prior estimates of the commonalities. Taking a step back, our original hypothesis was centered on the idea that we should have as many Factors as industry sectors. Our original sectors consisted of (1) Banking, (2) Oil Field Services, (3) Oil Refining, and (4) Industrial – Chemical. We currently have five factors in this model, but the fifth factor doesn't appear to provide us with much based on the values. Also, within each factor loadings, certain stocks are represented. Factor.1 has BHI, HAL, HES, and SLB. Factor.2 has BAC, JPM, and WFC. Factor.3 has DD, DOW, and HUN. Finally, Factor 4 has CVX and XOM. Factor.5 is the odd one out and doesn't appear to represent any market sector strongly. Based on the information we just presented, this model seems to be valid.

Figure 12: Goodness of Fit.



Goodness of Fit	
Chi-Square without Bartlett's Correction	11.098156
Akaike's Information Criterion	-20.901844
Schwarz's Bayesian Criterion	-88.367542
Tucker and Lewis's Reliability Coefficient	1.00584

The goodness of fit indices appears to indicate that this is a well-fitted model. Once again, the (AIC) measures the best number of parameters to include in a model when maximum likelihood estimation is used. Typically, the number of factors that yields the smallest value of AIC is considered best. The SBC gauges the best number of parameters and the number of factors that generate the smallest value of SBC is considered to be best. The Tucker and Lewis’s Reliability Coefficient is an indicator of reliability, with value ranges from 0 to 1. A larger value indicates better reliability.

Conclusion:

We examined four different factor analysis modeling techniques. In each case, we implemented different methodologies to arrive at the appropriate amount of factors. The use of Factor Rotation greatly improves the interpretability of the model, improves the simple factor structure and maintains the underlying mathematical properties. Incorporation of Maximum likelihood estimation gives us added ability to model interpretation and makes use of weighting correlations by their uniqueness. Maximum prior commonalities appear to have produced the model that best fits with our initial hypothesis while maintaining the necessary information needed to understand the appropriateness of the overall model.

Apendix:SAS Code

```
/*Read file into SAS and identify it as mydata*/
libname mydata "/scs/wtm926/" access=readonly;
/***1. In order to get better factor analysis results, let’s drop some of the variables from the data set and create the return data set. We will be left with: (1) Banking, (2) Oil Field Services, (3) Oil Refining, and (4) Industrial – Chemical.***/
data temp;
set mydata.stock_portfolio_data;
* Let's drop some variables to get better factor analysis results;
drop AA HON MMM DPS KO PEP MPC GS ;
run;
proc contents data=temp order=varnum;
run;
proc sort data=temp; by date; run; quit;
data temp;
set temp;
* Compute the log-returs;
* Note that the data needs to be sorted in the correct
direction in order for us to compute the correct return;
return_BAC = log(BAC/lag1(BAC));
return_BHI = log(BHI/lag1(BHI));
return_CVX = log(CVX/lag1(CVX));
return_DD = log(DD/lag1(DD));
return_DOW = log(DOW/lag1(DOW));
return_HAL = log(HAL/lag1(HAL));
return_HES = log(HES/lag1(HES));
return_HUN = log(HUN/lag1(HUN));
```

```
return_JPM = log(JPM/lag1(JPM));
return_SLB = log(SLB/lag1(SLB));
return_WFC = log(WFC/lag1(WFC));
return_XOM = log(XOM/lag1(XOM));
* Compute the remainder of the log-returs;
response_VV = log(VV/lag1(VV));
run;
proc print data=temp(obs=10); run; quit;
/***Next step***/
data return_data;
set temp (keep= return_:);
* What happens when I put this keep statement in the set statement?;
* Look it up in The Little SAS Book;
run;
proc print data=return_data(obs=10); run;
/*** 2. We will begin our factor analysis by performing a Principal Factor Analysis without a factor
rotation.
 Under this SAS procedure call SAS will automatically select the number of factors to retain.***/
ods graphics on;
proc factor data=return_data method=principal priors=smc rotate=none plots=(all);
run; quit;
ods graphics off;
/***3. Now let us apply a VARIMAX rotation to the Principal Factor Analysis***/
ods graphics on;
proc factor data=return_data method=principal priors=smc rotate=varimax plots=(all);
run; quit;
ods graphics off;
```

```
/*** 4. ) Now let us use (Maximum Likelihood Factor Analysis) with a VARIMAX rotation***/
ods graphics on;
proc factor data=return_data method=ML priors=smc rotate=varimax plots=(loadings);
run; quit;
ods graphics off;
/*** 5. Every factor analysis procedure requires estimates of the prior communalities
Let’s consider a Maximum Likelihood Factor Analysis with a VARIMAX rotation
but with the MAX argument for the PRIORS option***/
ods graphics on;
proc factor data=return_data method=ML priors=max rotate=varimax plots=(loadings);
run; quit;
ods graphics off;
quit;
```

References

(1) Black, K. (2008). Business statistics: For contemporary decision making. Hoboken, NJ: Wiley.

(2) Montgomery, D. C., Peck, E. A., Vinning, G. G., (2012). Introduction to Linear Regression Analysis Hoboken, NJ: Wiley.

(3) Everitt, B. C. (2010). Basic Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences. Sound Parkway, NW: CRC Press.

(4) Cody, R. (2011). SAS: Statistics by Example. Carey, NC: SAS Institute Inc.

(5)https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_factor_sect006.htm (accessed February 17, 2017)

(6) https://support.sas.com/documentation/cdl/en/statugfactor/61783/PDF/default/statugfactor.pdf (accessed February 17, 2017)

(7) http://www.utdallas.edu/~nkumar/FactorExample.PDF (accessed February 17, 2017)

(8) https://onlinecourses.science.psu.edu/stat505/node/82 (accessed February 17, 2017)

(9) http://www.jmp.com/support/help/The_Factor_Analysis_Report.shtml (accessed February 17, 2017)