Assignment #7

Scott M. Morgan

Introduction:

The purpose of this assignment is to use factor analysis to identify sectors in the stock market. To accomplish this objective, I use the stock portfolio data set provided by the course instructor and SAS Studio to perform factor analysis with and without rotation. The U.S. stock market is represented by the Vanguard Large-Cap ETF. Within the context of factor analysis, we posit that we have 3 or 4 factors (i.e. sectors) in the data set. The data set consists of the daily closing prices of 20 publicly traded companies between the 2-year period of January 3, 2012 to December to 31, 2013. Factor analysis is an analytical technique closely related to principal component analysis, the focus of Assignment #6; however, it is designed to identify latent variables that are represented by correlated response variables. Factor analysis is not only a useful tool for investors to predict sectors but, framed differently, is also a useful tool for decomposing what latent factors drive market or portfolio returns. These factors, which are beyond the scope of this assignment, include attributes like market timing, allocation, selection, style and industry exposure, to name a few. Similar to the previous exercise, I expect this assignment will be an excellent application of this technique as financial market analysis can easily become overwhelming given the volume, variety and velocity of the data involved.

Results:

In the subsequent sections, we prepare the stock portfolio data for analysis and perform factor analysis with and without factor rotation. Factor rotation is useful as it provides a means to obtain more interpretable factor loadings.

Data Preparation. In this section, we perform our initial data preparation. The raw data consists of daily closing stocks prices for 20 publicly traded U.S. equities and the Vanguard Large-Cap ETF (VV). To obtain better factor analysis results, we drop 8 variables from the data set (Tickers: AA, HON, MMM, DPS, KO, PEP, MPC and GS). By eliminating these companies, we are left with 12 publicly traded U.S. equities; specifically 3 within each of the Banking, Oil Field Services, Oil Refining and Industrial – Chemicals sectors. The VV is used as a proxy for the broad U.S. stock market as it currently holds 614 individual stocks and exhibits a trailing 5-year tracking error of 0.08% to its benchmark, the CRSP US Large Cap Index, according to Morningstar as of March 31, 2017. The CRSP US Large Cap Index includes U.S. companies that comprise the top 85% of investable market capitalization. It includes both mid and mega capitalization companies and is what I believe to be a relatively suitable proxy for the investable universe of the U.S. stock market. Tracking error, or active risk, is the annualized standard deviation of daily return differences between the return performance of the fund and the return performance of its underlying index. As ETFs are intended to mirror non-investable market indexes, we want this number as small as possible. For context, a trailing 5-year tracking error of 0.08% is considered exceptional so we are comfortable with the VV as a market proxy.

We calculate the return on the stocks as follows:

$$r_i = \frac{p_i - p_j}{p_j}$$

Where *r* is the return at time *i* and *p* is the closing price of the security at time *i* and *j*. We use return instead of price to measure all variables in a comparable metric. We also compute the natural log of returns. Log returns are preferred for financial analysis for reasons of normalization and they are usually not auto-correlated while prices can be (Aldridge, 2013). Analysis might also benefit from log transformations if there are a wide range of x values; which is the case with our sample of 20 different stocks and their prices over the 2-year period. This is the same methodology used to calculate the returns for the previous assignment.

We decide a priori that the criteria for significance of the factor loadings to be 0.5. Any loading above this threshold will be interpreted in the context of the factor(s) that have been identified. This cutoff is chosen given it is relatively conservative.

Principal Factor Analysis (PFA). We begin our factor analysis by performing a Principal Factor Analysis (PFA) without a factor rotation. Within the context of factor analysis, our hypothesis was that we have 3 or 4 factors (i.e. sectors), in the data set. Table 1 below provides the Eigenvalues of the reduced correlation matrix.

Table 1: Eigenvalues of the Reduced Correlation 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	-.05321803	0.01517115	-0.0078	1.1011
7	-.06838918	0.03291807	-0.0100	1.0911
8	-.10130725	0.01600696	-0.0148	1.0763
9	-.11731422	0.00866270	-0.0171	1.0593
10	-.12597692	0.01040221	-0.0184	1.0409
11	-.13637913	0.00786652	-0.0199	1.0210
12	-.14424565		-0.0210	1.0000

The first 2 eigenvalues account for most variance; with the first explaining much more than the second. This suggests the variables are highly-correlated and there is a latent trait among the variables. None of the communalities are close to 1.0 so there is no indication of an issue in that regard. The scree plot below (Exhibit 1) confirms that the first Eigenvalue explains much of the variation. There is also a break in slope between Eigenvalues 2 through 4 which could be interesting but thereafter appears insignificant. One strange observation about the SAS output is that the cumulative variance explained is greater than 100% across the majority of Eigenvalues. The SAS procedure used automatically selects the

number of factors to retain, which in this case is 2 (Table 3 below). SAS used the proportion criterion to select the number of factors to retain. The default value is 1.0 or 100%. Due to our removal of the aforementioned 8 stocks we expected there would be 3 to 4 sectors (or factors) identified; so the initial hypothesis is not supported and it is surprising to see only 2 factors retained. This is evidence that there might be some other common element between the stocks aside from sector that is giving rise to the high correlation between these variables.

Exhibit 1: Scree Plot

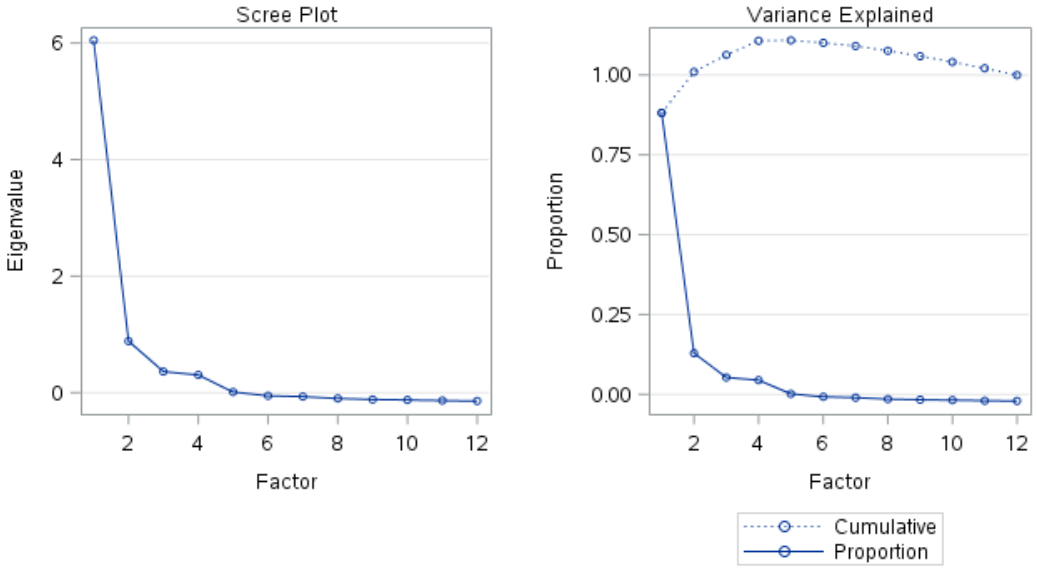

Table 2 below provides the factor pattern matrix of the 2 factors. Examining the loadings of the factor patterns we see that all variables are highly correlated with Factor 1 according to the 0.5 cutoff established at the onset of the exercise so we do obtain a simple structure. None of the variables are highly correlated with Factor 2.

Table 2: Factor Pattern Matrix

	Factor 1	Factor 2
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

Table 3: Variance Explained by Each Factor

Factor 1	Factor 2
6.0473258	0.8847081

Exhibit 2: Factor Pattern

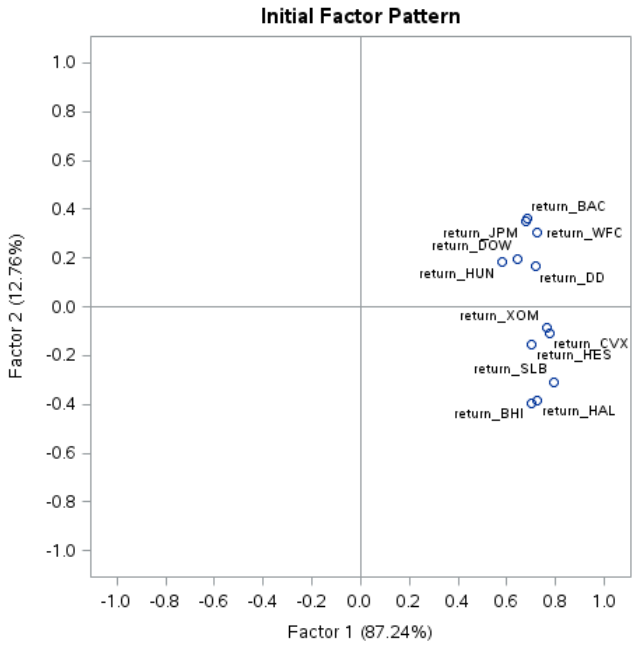

Exhibit 2 above also shows a distinct pooling along the x-axis related to Factor 1. Based on these diagnostics, we posit the driving force behind the relationships in Factor 1 are related to systematic market factors or movements. This idea is based on the capital asset pricing model (CAPM) which describes the relationship between systematic and unsystematic risk. The CAPM is a widely used and popular theory in capital market literature. We attribute the meaning of Factor 2 to sector membership or some firm idiosyncratic characteristic.

Principal Factor Analysis (PFA) with VARIMAX Rotation. We now perform a PFA with a factor rotation. Factor rotation is intended to make interpretation of factors easier and more reliable (Thurstone, 1931). We first use the VARIMAX rotation technique, an orthogonal rotation that maximizes the squared factor loadings in each factor. The reduced correlation matrix is identical to the above, so we only provide the updated factor patterns in Table 4 below.

Table 4: Rotated Factor Pattern Matrix

	Factor 1	Factor 2
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

Table 5: Variance Explained by Each Factor

Factor 1	Factor 2
3.4711423	3.4608916

None of the communalities are close to 1.0 so there is no indication of an issue in that regard. We observe that SAS retains the same number of factors (Table 5) but the rotation has changed several components of the output. Specially, based on the 0.5 loading limit, Factor 1 now consists of BAC, DD, DOW, HUN, JPM and WFC. Factor 2 now consists of BHI, CVX, HAL, HES, SLB and XOM. The interpretability of the analysis is easier with the rotation and we do obtain a 'simple structure' as each variable loads highly into one and only one factor. Table 6 below summarizes the companies in the analysis with the VARIMAX rotation by sector and factor. We can clearly state that Factor 1 is comprised of Banking and Industrial – Chemical firms while Factor 2 is comprised of Oil Field Services and Oil Refining firms.

Table 6: Summary of Principal Factor Analysis with VARIMAX Rotation

Ticker	Sector	Factor
BAC	Banking	Factor 1
JPM	Banking	Factor 1
WFC	Banking	Factor 1
DD	Industrial - Chemical	Factor 1
DOW	Industrial - Chemical	Factor 1
HUN	Industrial - Chemical	Factor 1
BHI	Oil Field Services	Factor 2
HAL	Oil Field Services	Factor 2
SLB	Oil Field Services	Factor 2
CVX	Oil Refining	Factor 2
HES	Oil Refining	Factor 2
XOM	Oil Refining	Factor 2

Maximum Likelihood (ML) Factor Analysis with VARIMAX Rotation. We now use Maximum Likelihood (ML) Estimation to estimate the common factors with a VARIMAX rotation. The Maximum Likelihood Estimation is regarded as the most respectable method of estimating parameters in the factor analysis model (Everitt, 2009). This method defines a type of "distance" between the observed covariance matrix and the covariance implied by the factor analysis model. We examine the factor patterns in Table 7 below.

Table 7: Rotated Factor Pattern Matrix

	Factor 1	Factor 2
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

Table 8: Variance Explained by Each Factor

	Weighted	Unweighted
Factor1	8.7156851	3.55022275
Factor2	10.1803287	3.42320994

From Table 8, we can see that 2 common factors are suggested by ML Factor Analysis with VARIMAX rotation. The model arrives at the number of factors by using the proportion criterion. None of the communalities are close to 1.0 so there is no indication of an issue in that regard. The factor loadings patterns are identical to the PFA so there is no difference in interpretability. From a modeling perspective, the advantages of the MF Factor Analysis include criteria for goodness-of-fit and it is transportable to other models for comparison.

Maximum Likelihood (ML) Factor Analysis with VARIMAX Rotation and MAX Priors. The next model considers a ML Factor Analysis with a VARIMAX rotation but with the MAX argument for the PRIORS options. Setting priors parameter to MAX sets the prior communality estimate for each variable to its maximum absolute correlation with any other variable. We should expect the cutoff for accepting factors to be different. We examine the factor patterns in Table 9 below.

Table 9: Variance Explained by Each Factor

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
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	0.26440	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

Table 10: Variance Explained by Each Factor

	Weighted	Unweighted
Factor 1	9.48177257	2.55119512
Factor 2	6.95572063	2.08400430
Factor 3	5.26449075	1.82173920
Factor 4	5.80237050	1.59069819
Factor 5	0.31984016	0.12246466

From Tables 9 and 10, we can see that 5 common factors are suggested by ML Factor Analysis. ML Factor Analysis with VARIMAX Rotation and Max Priors arrives at the number of factors by using the proportion criterion. None of the communalities are close to 1.0 so there is no indication of an issue in that regard. As hypothesized, there are 4 sectors in the data set. Specifically, Factor 1 indicates Oil Field Services, Factor 2 indicates Banking, Factor 3 indicates Industrial – Chemical and Factor 4 indicates Oil Refining. Interestingly, Factor 5 does not contain any significant loadings nor does HES meet any of the loading thresholds. We would expect that HES be included in Factor 4 (Oil Refining). While this doesn't render the analysis invalid, it may be advantageous to limit the number of factors to 4. The output in Tables 9 and 10 suggest that factor selection is highly dependent on the prior estimates of communalities.

Maximum Likelihood (ML) Factor Analysis with VARIMAX Rotation, Max Priors and NFACTORS = 4. To a supplement the required models in the exercise instructions, we run a final model restricting the number of factors to 4. The output is presented in Tables 11 and 12 below.

Table 11: Variance Explained by Each Factor

	Factor 1	Factor 2	Factor 3	Factor 4
return_BAC	0.19757	0.75423	0.27205	0.17115
return_BHI	0.75560	0.14546	0.19113	0.24328
return_CVX	0.37940	0.24873	0.26653	0.71355
return_DD	0.24525	0.27169	0.65220	0.30797
return_DOW	0.19069	0.24648	0.67301	0.22569
return_HAL	0.81890	0.18590	0.21348	0.16413
return_HES	0.48670	0.24951	0.25485	0.39568
return_HUN	0.23250	0.27340	0.58577	0.07364
return_JPM	0.20509	0.76170	0.23935	0.17940
return_SLB	0.72981	0.25515	0.24982	0.29668
return_WFC	0.20922	0.60742	0.36668	0.29113
return_XOM	0.37625	0.29148	0.25235	0.65233

Table 12: Variance Explained by Each Factor

	Weighted	Unweighted
Factor1	9.39289839	2.57196785
Factor2	6.69316710	2.05723775
Factor3	5.10330088	1.83137584
Factor4	5.71876627	1.56282992
Factor1	9.39289839	2.57196785

With the elimination of Factor 5, the loadings change but the results are not starkly different. One observation is the loading value increases for the HES variable. In order to associate HES with a factor, however, we have to ease our 0.50 threshold slightly to 0.45. If this were the case, HES would be associated with Factor 1 (Oil Field Services) instead of Factor 3; where other Oil Refining firms, its original classification, are grouped. This could be a result of the company trading or being valued more like an Oil Field Service firm because of certain business lines, revenue exposures, client demographics or other esoteric characteristics. We conclude that this supplementary model with the limited number of factors has the cleanest factor structure – item loadings above 0.45 (~0.50), no crossloadings, simple structure – and therefore the best fit to the data.

Reference(s):

Everitt, B. (2009). *Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences*. Boca Raton, FL: CRC Press.

Thurstone, L.L. (1931). *Mulitple factor analysis.* Psychological Review, 39, 406-427.

Code:

```
libname mydata "/scs/wtm926/" access=readonly;



*************************************************************************************
*******;
* Part 1 - Data Import and Preparation;
*************************************************************************************
*******;



data temp;
        set mydata.stock_portfolio_data;
run;


proc sort data=temp; by date; run; quit;


* 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;


data temp;
 set temp;
 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));
 *return_VV  = log(VV/lag1(VV));
 response_VV = log(VV/lag1(VV));
run;

proc print data=temp(obs=10); run; quit;

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;



*************************************************************************************
*******;
* Part 2 - Principal Factor Analysis Without Factor Rotation;
*************************************************************************************
*******;
```

```
ods graphics on;

proc factor data=return_data
        method=principal
        priors=smc
        rotate=none
        plots=(all);
run; quit;

ods graphics off;



*************************************************************************************
*******;
* Part 3 - Principal Factor Analysis With VARIMAX Rotation;
*************************************************************************************
*******;



ods graphics on;

proc factor data=return_data
        method=principal
        priors=smc
```

```
        rotate=varimax
        plots=(all);
run; quit;

ods graphics off;

*********************************************************************************
*******;
* Part 4 - Maximum Likelihood Factor Analysis with VARIMAX Rotation
*********************************************************************************
*******;

ods graphics on;

proc factor data=return_data
        method=ML
        priors=smc
        rotate=varimax
        plots=(loadings);
run; quit;

ods graphics off;
```

```
*************************************************************************************
*******;
* Part 5 - Maximum Likelihood Factor Analysis with VARIMAX Rotation and MAX PRIORS
*************************************************************************************
*******;



ods graphics on;


proc factor data=return_data method=ML priors=max rotate=varimax
  plots=(loadings);
run; quit;


ods graphics off;



*************************************************************************************
*******;
* Part 6 - Supplmentary - Maximum Likelihood Factor Analysis with VARIMAX Rotation and MAX PRIORS
- NFACTOR = 4
*************************************************************************************
*******;



ods graphics on;


proc factor data=return_data method=ML priors=max rotate=varimax
```

```
  plots=(loadings scree)
  nfactors = 4;
run; quit;


ods graphics off;
```