Assignment #7

Ji Jung

**Introduction**

For this assignment, the factor analysis will be used to identify sectors in the stock market data. For better results in factor analysis, values may be dropped from the data set. With the elimination, there will be four sectors remaining; Banking, Oil Field Services, Oil Refining, and Industrial - Chemical. Within the context of factor analysis, a hypothesis can be constructed that there are three or four factors (or industry sectors) within this data set.

**Results**

**Principal Factor Analysis**

First step is to perform a Principal Factor Analysis without a factor rotation. The SAS procedure will automatically select the number of factors to retain. Factor analysis begins by substituting the diagonal of the correlation matrix with prior communality estimates. The communality estimate for a variable is the estimate of the proportion of the variance of the variable that is both error free and shared with other variables within the matrix. This calculation is completed by using the SMC method, which uses the squared multiple correlation between the variable and all other variables. The observed Prior Communality Estimates reveal that there are some values close to one (greater than 0.6). It is uncertain if the SMC method will be the most appropriate for the modeling. Thus, examining the eigenvalues of the reduced correlation matrix:

| Observation | 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 |
| Observation | Eigenvalue | Difference | Proportion | Cumulative |
| 12 | -.14424565 | - | -0.0210 | 1.0000 |

Table 1: Eigenvalues of the Reduced Correlation Matrix

From Table 1, the first two eigenvalues have a very large proportion of the variance, specially the first has much more than the second. It may be an evidence that the variables within the model are all highly correlated with each other and that there is some latent quality or trait which may suggest high correlation amongst the variables. Factor analysis is utilized to retain insights into the quality or trait effecting the correlation. It is worth noting that a cumulative value that is greater than 1.0 for the first two eigenvalues. This may be due to the SMC method. Next, the loadings of the factor pattern and the respective factor variance is examined:

| | 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 2: Factor Pattern

| Factor 1 | Factor 2 |
|---|---|
| 6.0473258 | 0.8847081 |

Table 3: Variance Explained by Each Factor

Results of SAS have retained two factors under its default settings. The MINEIGEN parameter is not specified with the FACTOR statement. Since the SMC method is used, the SAS manual lists that the MINEIGEN will be calculated as:

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

Solving the formula results in $6\frac{86244298}{12}$. = 0.57. Given this calculation, it is expected to see two factors. Given the second factors sign, two groups can be differentiated. The first appears to subsume both the Banking and industrial sectors (BAC, DD, DOW, HUN, JPM, WFC), and the second appears to subsume the Oil refining and field services sectors (BHI, CVX, HAL, HES, SLB, XOM).

It seems that all the variables are highly loaded for the first factor, whereas the variables for the second factor do not meet the specified criteria for loading. If the results are interpreted systematically, the following equation for each variable within analysis can be made.

$X_1 = \lambda 1 f_1 + \lambda 2 f_2 + ,... + \lambda k f_k + u_1$

Using this formula, with BAC would return:

return_BAC = $0.68475 \times f_1 + 0.36021 \times f_2$

If strict loading criteria is called for, the second factor and loading coefficient within the equation would not be included. It may be practical to choose the dominant factor and state that the variable is explained more by the dominant factor as return_BAC is explained more by factor 1 than factor 2.

Two factors graphically via the graph is exampled below.

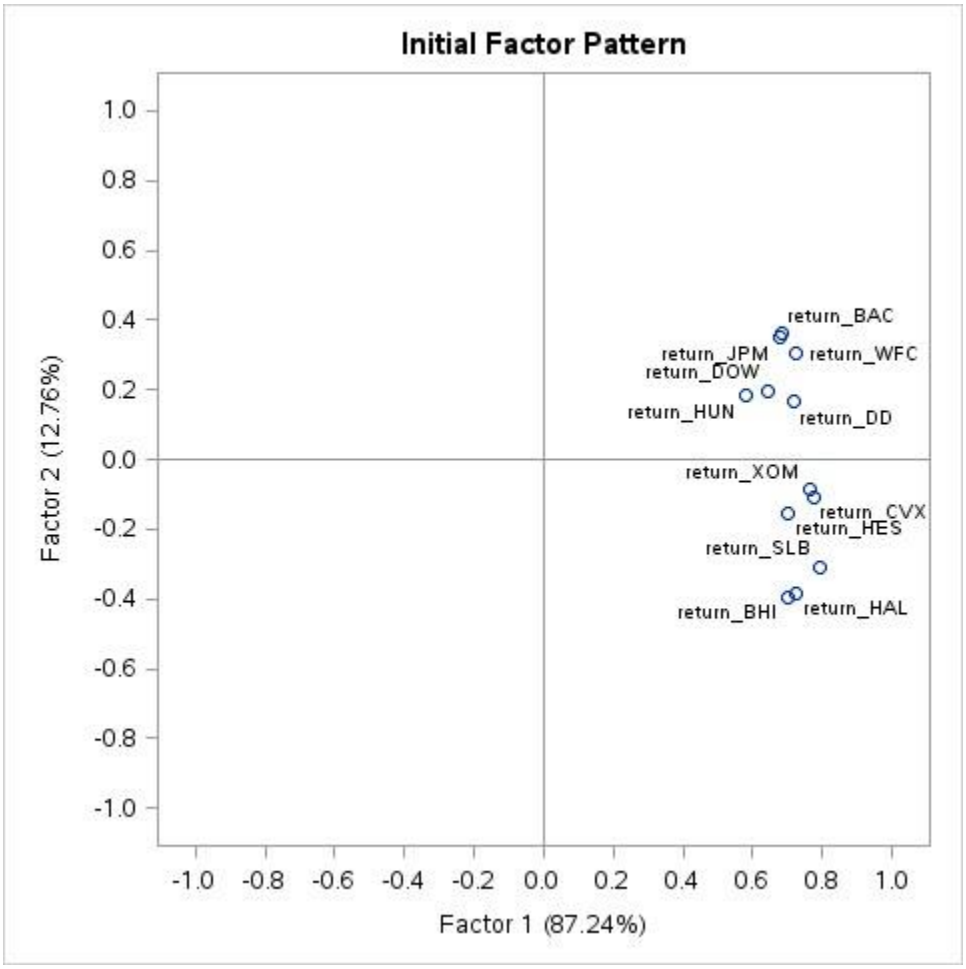


Figure 1: Initial Factor Pattern

With some creativity, the process of reification can be begun. The meaning of the first factor is to be representative of the overall market and would give it the name 'market'. For the second factor, it would attribute meaning of sector differentiation, and thus would give it the name 'sector'. The first

factor has variables which pass the loading threshold even though the second factor does not. This may not provide much interpretations thus further look at the sign within the second factor is needed.

**Principal Factor Analysis with Rotation (Varimax)**

A factor analysis with rotation is performed in this section. The rotation is meant to improve the interpretability of the model. By default, the unrotated output maximizes the variance accounted for by the first and subsequent factors, and forces the factors to be orthogonal. Rotation serves to make the output more understandable by seeking a simple structure, or a pattern of loadings where items load most strongly on one factor, and much more weakly on the other factors. The Varimax rotation is used as an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all variables (rows) in a factor matrix. The factor analysis outputs are the same as above, with new outputs for the rotation method:

| | 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 4: Rotated Factor Pattern

| Factor 1 | Factor 2 |
|---|---|
| 3.4711423 | 3.4608916 |

Table 5: Variance Explained by Each Factor

The interpretability is different for this model. The rotation allows to consider each factor as providing close to the same explanatory value for the variance within the model. Interpreting with the above decided 0.5 loading threshold allows us to see that Factor 1 is comprised of BAC, DD, DOW, HUN, JPM, WFC. Factor 2 is comprised of BHI, CVX, HAL, HES, SLB, XOM. Reification of the factors is much easier in

this model. It can be stated that Factor 1 is comprised of Banking and industrial sectors, whereas Factor 2 is comprised of Oil refining and field services sectors.

However, it is not transparent that a loaded factor for a single variable. If it were so, it would be proper to drop the variable from the model and consider it independently from the factor analysis.

**Maximum Likelihood Factor Analysis with Rotation (Varimax)**

A maximum likelihood factor analysis with varimax rotation will be performed next. This approach requires several assumptions which have not yet been validated throughout this course. The benefit in taking on these assumptions is that maximum likelihood is a formal estimation procedure which provides formal inferences for factor loadings and goodness-of-fit criteria. In observation, the method computes an initial set of eigenvalues to assess the convergence criterion. As with above, calculating the MINEIGEN default to be $\frac{18.8960127}{12}$ = 1.574667725. It means that a model would have two factors. Once the criterion is satisfied, there is two separate statistical hypothesis test with the null hypothesis stated as no common factors and 2 factors are sufficient respectively. Both tests allow to accept the null hypothesis. The rotated factor pattern will be examined directly.

| | 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 6: Rotated Factor Pattern

| Factor | Weighted | Unweighted |
|---|---|---|
| Factor1 | 8.7156851 | 3.55022275 |
| Factor2 | 10.1803287 | 3.42320994 |

Table 7: Variance Explained by Each Factor

The same number of common factors are suggested by the maximum likelihood method. The factor loadings between principal factor analysis and maximum likelihood with rotations are very similar, leaving no difference in interpretability. The added benefit from a maximum likelihood factor analysis is the goodness-of-fit criteria. The maximum likelihood methodology allows some criterion for model comparison.

**Maximum Likelihood Factor Analysis, with Rotation and Max Priors**

With the Max priors parameter set, it is likely that a drastically different threshold for accepting factors will be noticed. Max sets the prior communality estimate for each variable to its maximum absolute correlation with any other variable. Calculation of MINEIGEN default returns $\frac{27.8241868}{12}$ = 2.318682233. From the manual, that only two factors come out of this method, but instead five are noticed:

| | 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 8: Rotated Factor Pattern

| Factor | Weighted | Unweighted |
|---|---|---|
| Factor1 | 9.48177257 | 2.55119512 |
| Factor2 | 6.95572063 | 2.08400430 |
| Factor3 | 5.26449075 | 1.82173920 |
| Factor4 | 5.80237050 | 1.59069819 |
| Factor5 | 0.31984016 | 0.12246466 |

Table 9: Variance Explained by Each Factor

This suggests that the factor selection is highly dependent on the prior estimates of communalities. The Max priors procedure seems to be a highly inclusive method for computing communalities. From the rotated factor patterns, it can be assumed that some of the factors will only be inclusive of a small subset of variables, based on loading conditions. If closer examination, it can be found that:

Factor 1: BHI, HAL, SLB

Factor 2: BAC, JPM, WFC Factor 3: DD, DOW, HUN Factor 4: CVX, XOM

Factor 5: No selections based on loading criterion

From previous assignment, these organizations correspond to the following:

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

Table 10: Ticker and Sector

Seemingly, expected results of factor lines drawn by sector appears to be occurring within this model.

Factor 1 strongly indicates Oil Field services, Factor 2 strongly indicates Banking, Factor 3 strongly indicates Industrial - Chemical, Factor 4 strongly indicates Oil Refining. This leaves out HES, which appears to load more heavily in the Oil Field services Factor than Factor 4 based on its sector. Unexpectedly, there is a fifth factor which has no a single variable within, based on loading criterion.

In conclusion, the method chosen for priors calculation and communalities is highly influential over the chosen factors from the model. Using Max returned closer results to initially expected from the familiarity with the data set. Nonetheless, the fifth factor has not yet provided any explanatory utility and seems an aberration of the model.