Factor Analysis of Weekly Trends in the Dow Jones Industrial Index: Predicting Future Stock Returns

A M Weerakkody

S/18/834

Contents

1 Introduction

Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) are widely utilized statistical techniques in the field of psychometrics and social sciences to identify hidden patterns of variables and to validate models. EFA aims to explore the basic structures of observed variables by identifying common factors that explain their correlations, thereby developing hypotheses and theoretical frameworks. On the other hand, CFA verifies or dismisses a proposed factor model, enabling a thorough evaluation of how well our theoretical model fits the observed data. In this mini-project, through practical demonstrations and analysis, we aim to gain insights into the complex relationships between variables and their hidden concepts, helping to gain a deeper understanding of complex phenomena within our chosen domain.

2 Methodology

For this study, I have chosen this Dow Jones Index dataset which consist of 750 columns and 16 variables. This dataset contains weekly data for the Dow Jones Industrial Index. It has been used in computational investing research. In this dataset, each record (row) is data for a week. Each record also has the percentage of return that stock has in the following week (percent_change_next_weeks_price). Ideally, this could be used to determine which stock will produce the greatest rate of return in the following week.

- **quarter:** Yearly quarter (1: Jan-Mar; 2: Apr-Jun).
- **stock**: stock symbol
- **date**: Last business day of the work (this is typically a Friday)
- **open**: Price of the stock at the beginning of the week
- **high:** Highest price of the stock during the week
- **low:** Lowest price of the stock during the week
- **close:** Price of the stock at the end of the week
- **volume:** Number of shares of stock that traded hands in the week
- **PCP:** Percentage change in price throughout the week
- **PCVLW:** Percentage change in the number of shares of stock that traded hands for this week compared to the previous week
- **PWV:** Number of shares of stock that traded hands in the previous week
- **NOW:** Opening price of the stock in the following week
- **NWC:** Closing price of the stock in the following week
- **PCNWP**: Percentage change in price of the stock in the
- **DD:** Number of days until the next dividend
- **PRND:** Percentage of return on the next dividend

Since this data set contains 15 columns, analyses and interpret process is not very easy. Therefore dimension reduction is required. Thus, the purpose of this analysis is to perform Explanatory Factor Analysis and Confirmatory Factor Analysis on this Dow Jones Index dataset. Here we are using exploratory factor analysis techniques such as Eigen values and Eigen vectors, factor loadings, communalities. Also we are focus on confirmatory factor model with some latent variables and corresponding graphs in this analysis.

3 Results and Conclusion

3.1 Exploratory Factor Analysis

3.1.1 Adequacy Test

3.1.1.1 KMO Test

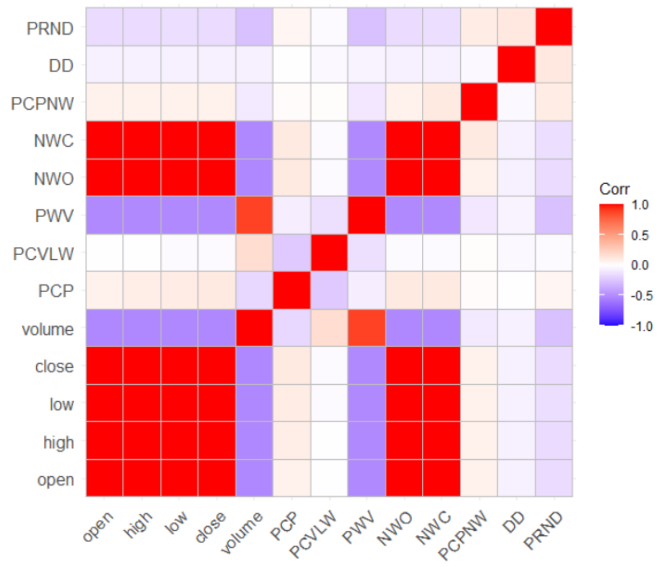
```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = normalized_data)
Overall MSA =  0.81
MSA for each item =
  open   high    low  close volume    PCP  PCVLW    PWV    NWO    NWC  PCPNW
  0.86   0.92   0.95   0.86   0.67   0.19   0.11   0.67   0.87   0.87   0.07
PRND     DD
0.82    0.80
```

Since the overall values for KMO test is 0.81, we can say our selected dataset is highly adequate for factor analysis.

3.1.1.2 Bartlett's Test

Test	DF	Chi-Squared	p-value
H0 : Correlation Matrix is an identity matrix Vs. H1 : Correlation matrix is not an identity matrix	78	2989.04	0

Since the p-value for Bartlett's Test is 0, we can say our selected dataset is highly adequate for factor analysis.



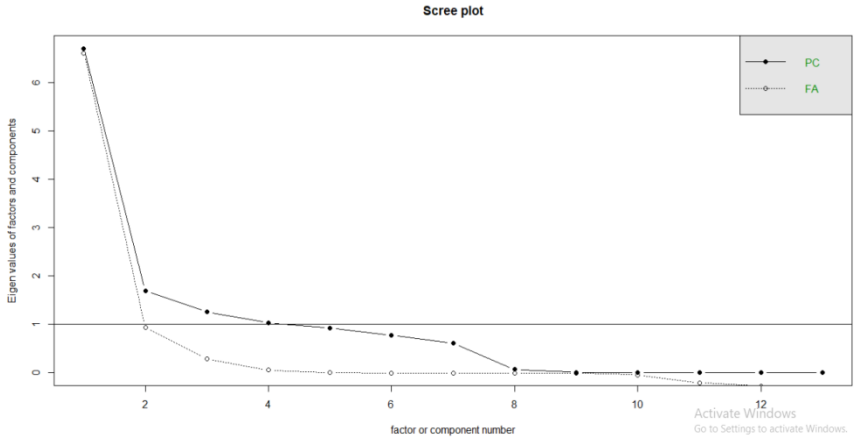
3.1.2 Correlation

In this above corrplot we can see some Higher correlation as well as very low correlations among the variables.

3.1.3 Eigen Values and variance of each component

Component	Eigen Value	Proportion of Variance	Cumulative Proportion
1	6.7001	0.5154	0.5154
2	1.6825	0.1294	0.6448
3	1.248	0.096	0.7408
4	1.0228	0.0787	0.8195
5	0.9188	0.0707	0.8902
6	0.7642	0.0588	0.949
7	0.6014	0.0463	0.9953
8	0.0603	0.0046	0.9999
9	0.0009	0.0001	1
10	0.0005	0	1
11	0.0003	0	1
12	0.0002	0	1
13	0.0001	0	1

The summation of the Eigen values show the total variance of the standardized variables which also equal to the number of variables 13. We have to determine the number of factors before starting the analysis. The first four components (where eigen values >1) explain 81.95% of the total variance which is a sufficient interpretation for the data set with a slight loss of information.



However since the 5th eigen values is very close to 1, and since including it increase the percentage of variance explained to 89%, future analysis will be done based on 5 components.

3.1.4 Factor Analysis

3.1.4.1 Exploratory Factor Analysis with Principle Component Analysis

3.1.4.1.2 Factor Loadings

	Factor1	Factor2	Factor3	Factor4	Factor5
open	1	0	0.01	0.02	0.01
high	1	0	0.03	0.02	0.01
low	1	0	0.03	0.02	0
close	1	0	0.05	0.02	0
volume	-52	0.81	-0.09	-0.03	0.15
PCP	0.07	-0.05	0.85	0.01	-0.15
PCVLW	-0.01	0.05	-0.14	0.01	0.78
PWV	-0.51	0.8	-0.04	-0.04	-0.23
NWO	1	0	0.05	0.02	0.01
NWC	1	0	0.05	0.06	0
PCPNW	0.05	-0.05	0.01	0.88	0.01
DD	-0.06	-0.13	-0.02	-0.03	-0.04
PRND	-0.15	-0.43	0.04	0.1	0.01

3.1.4.2 Exploratory Factor Analysis with Maximum Likelihood Method

3.1.4.2.1. Factor Loadings

	ML1	ML3	ML2	ML4	ML5
open	1	0.02	0.01	0.01	-0.01
high	1	0.02	0.01	0.03	0
low	1	0.02	0	0.03	-0.01
close	1	0.02	0	0.04	-0.01
	-0.5	-0.03	0.12	-0.08	0.81
PCP	0.08	0.01	-0.12	0.99	-0.05
PCVLW	0.04	0.01	0.99	-0.12	0.05
PWV	-0.51	-0.04	-0.18	-0.02	0.8
NW	1	0.02	0	0.04	-0.01
NWC	1	0.06	0	0.05	-0.01
PCPNW	0.05	1	0.01	0.01	-0.05
DD	-0.06	-0.03	-0.03	-0.01	-0.11
PRND	-0.15	0.09	0.01	0.04	-0.43

3.1.5 Communalities

	PCA Method	ML Method
NWC	1	0.9984
close	0.9998	0.9989
NWO	0.9997	0.9988
high	0.9995	0.9988
low	0.9986	0.9986
open	0.9981	0.9982
volume	0.947	0.9522
PWV	0.964	0.9522
PCPNW	0.7328	0.995
PCVLW	0.651	0.7438
PRND	0.222	0.2128
PCP	0.0999	0.0998
DD	0.0218	0.017

- PCA model explains NWC, close, NOW, high, low, open, volume, PWV the best and is not bad for PCPNW and PCVLW.
- ML model explains NWC, close, NOW, high, low, open, volume, PWV as well as PCPNW the best and is not bad for PCPNW also.
- However, for PCP and DD variables both model do not do a good job, explaining a very low variances.

3.1.6 Comparing PCA model and ML model

PCA Model

	PC1	PC2	PC4	PC3	PC5
SS Loadings	6.55	1.51	0.79	0.76	0.71
Proportion of Variance	0.4	0.12	0.06	0.06	0.05
Cumulative Variance	0.5	0.62	0.68	0.74	0.79
Proportion Explained	64	0.15	0.08	0.07	0.07
Cumulative Proportion	0.64	0.78	0.86	0.93	1

```
The harmonic n.obs is 720 with the empirical chi square 5.91  with prob < 1
The total n.obs was 720 with Likelihood Chi Square =  1058.17  with prob <
5.6e-209

Tucker Lewis Index of factoring reliability = 0.882
RMSEA index = 0.25
```

ML Model

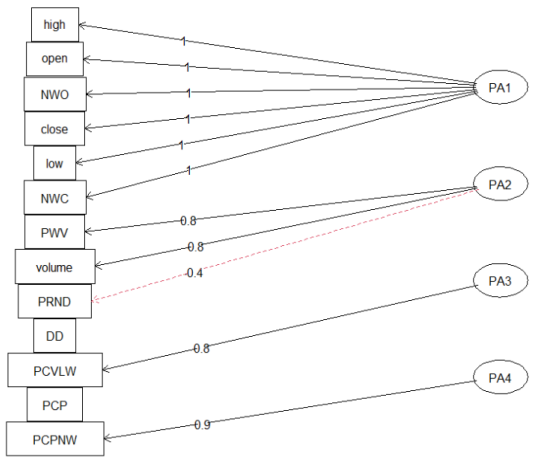
	ML1	ML5	ML2	ML3	ML4
SS Loadings	6.53	1.52	1.04	1.01	1.01
Proportion Variance	0.5	0.12	0.08	0.08	0.08
Cumulative Variance	0.5	0.62	0.7	0.78	0.85
Proportion Explained	0.59	0.13	0.09	0.09	0.09
Cumulative Proportion	0.59	0.72	0.82	0.91	1

```
The harmonic n.obs is 720 with the empirical chi square 6.59  with prob <  1
The total n.obs was 720 with Likelihood Chi Square =  2637.22  with prob <  0


Tucker Lewis Index of factoring reliability = 0.701
RMSEA index = 0.39
```

Both models are statistically significant models.

Even though the cumulative variance explained by ML model is high than the PCA model, Since the Tucker Lewis Index of ML model is lower than the PCA model and since RMSEA index of ML model is higher than PCA model, we can more rely on PCA model than ML model. Therefore we can conclude that for this dataset PCA model gives a better Factor Analysis than ML method.

Factor Diagram



3.1.7 Factor Model

- **Factor 1** : Factor 1 is strongly positively correlated with high, open, close, low, next_week_open (NWO), next_week_close (NWC). We can infer that Factor 1 likely represents a factor related to the overall movement or performance of the stock market or a specific segment of it. A common name for factor 1 could be : "Market Sentiment Factor".
- **Factor 2** : Factor 2 has strong positive correlations with previous_week_volume (PWV) and volume. This suggests that Factor 2 is associated with trading volume. In addition Factor 2 has a moderate negative correlation with percent_return_next_dividend (PRND). It might indicate that stocks with higher trading volumes tend to have lower dividend returns. A common name for Factor 2 could be: "Trading Activity Factor".
- **Factor 3:** Factor 3 only has a significant correlation with percent_change_volume_over_last_week (PCVLW). This suggests that Factor 3 is primarily associated with changes in trading volume over the last week. a common name for Factor 3 could be: "Volume Change Factor".
- **Factor 4 :** Factor 4 only has a significant correlation with the percentage change in price of the stock (PCPNW). This suggests that Factor 4 is primarily associated with changes in the price of the stock for the next week. A common name for Factor 4 could be: "Price Change Factor"

3.2 Confirmatory Factor Analysis

Confirmatory Factor Analysis Model

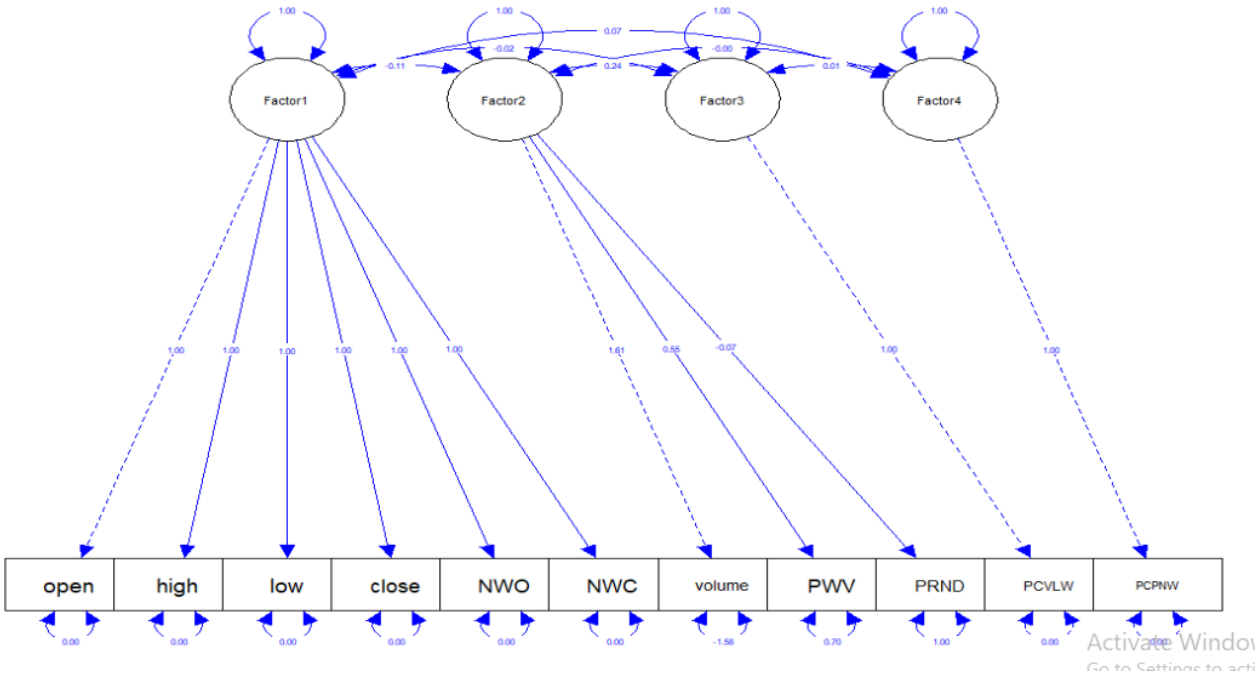
```
Model Test User Model:

  Test statistic                2365.564
  Degrees of freedom              40
  P-value (Chi-square)            0.000

Model Test Baseline Model:

  Test statistic                29201.452
  Degrees of freedom              55
  P-value                         0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)     0.920
  Tucker-Lewis Index (TLI)        0.890
```

I have predefined 3 factors using these variables.

Factor1 =~ open + high + low + close + NWO + NWC
Factor2 =~ volume + PWV + PRND
Factor3 =~ PCVLW
Factor4 =~ PCPNW

The output looks like this.

Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) assess the fit of the model compared to the Baseline model. Values closer to 1 indicate better fit. In this case, CFI is 0.920 and TLI is 0.890, suggesting acceptable model fit.



4 Conclusion And Recommendation

- According to the analysis we can reduce our 13 variable dataset to 4 Latent factors. I proved from empirical chi-squared test 4 factors are sufficient to describe the dataset. However four factor model explains only 79% of variance of the dataset.
- We could observe very high correlations among variables
- According to the results of factor loadings using PC method, Factor1 strongly correlated with 6 variables. Only the 'PRND' variable has moderate negative correlation with Factor2. Other 2 factors only correlate with one single variable of the dataset.
- Without using any factor rotation technique, Factor loadings are not giving any clear conclusion about the model.
- Both PC model and FA models explains most of the variables well,only one two variables were not able to explained.
- In CFA the chi-square test suggests that the model does not fit the data well, but CFI & TLI indicates that the model is an acceptable fit.
- In conclusion, while the model may not fit perfectly, it offers an acceptable representation of the underlying structure of the data according to several fit indices. Further refinement and exploration may help enhance the model's performance and better capture the complexities of the observed variables and latent constructs.

5 References

https://stats.oarc.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/

https://stats.oarc.ucla.edu/r/seminars/rcfa/

https://pages.mtu.edu/~shanem/psy5220/daily/Day22/cfa.html#:~:text=Confirmatory%20factor%20analysis%20typically%20identifies,also%20useful%20in%20complex%20situations.

https://www.analysisinn.com/post/kmo-and-bartlett-s-test-of-sphericity/

https://statisticsbyjim.com/basics/factor-analysis/

https://statisticsbyjim.com/basics/factor-analysis/

https://www.scribd.com/document/441139421/Sample-Report-for-CFA-in-APA

6 Appendices

Data Set :
https://code.datasciencedojo.com/datasciencedojo/datasets/tree/master/Dow%20Jones%20Index?__hstc=148453154.03f05612f5d80639690e91edf8532a31.1712166063995.1712252152316.1712301692817.5&__hssc=148453154.1.1712301692817&__hsfp=463517827

Libraries :library(tidyverse) library(corrplot) library(psych) library(lavaan) library(ggplot2) library(factoextra) library(semPlot) library(ggcorrplot)

Loading Data :dow_jones_data <- read_csv("../Data/dow_jones_index.csv")

	quarter	stock	date	open	high	low	close	volume	PCP	PCVLW	PWV	NWO	NWC	PCPNW	DD	PRND
1	1	AA	2011-01-14	16.71	16.71	15.64	15.97	242963398	-4.4284900	1.3802230	239655616	16.19	15.79	-2.4706600	19	0.1878520
2	1	AA	2011-01-21	16.19	16.38	15.60	15.79	138428495	-2.4706600	-43.0249593	242963398	15.87	16.13	1.6383100	12	0.1899940
3	1	AA	2011-01-28	15.87	16.63	15.82	16.13	151379173	1.6383100	9.3555001	138428495	16.18	17.14	5.9332500	5	0.1859890
4	1	AA	2011-02-04	16.18	17.39	16.18	17.14	154387761	5.9332500	1.9874517	151379173	17.33	17.37	0.2308140	97	0.1750290
5	1	AA	2011-02-11	17.33	17.48	16.97	17.37	114691279	0.2308140	-25.7121949	154387761	17.39	17.28	-0.6325470	90	0.1727120
6	1	AA	2011-02-18	17.39	17.68	17.28	17.28	80023895	-0.6325470	-30.2266958	114691279	16.98	16.68	-1.7667800	83	0.1736110
7	1	AA	2011-02-25	16.98	17.15	15.96	16.68	132981863	-1.7667800	66.1776936	80023895	16.81	16.58	-1.3682300	76	0.1798560
8	1	AA	2011-03-04	16.81	16.94	16.13	16.58	109493077	-1.3682300	-17.6631500	132981863	16.58	16.03	-3.3172500	69	0.1809410

Remove Missing Values : na.omit(dow_jones_data)

Rename Column Names : new_names <- c("quarter" = "quarter", "stock" = "stock", "date" = "date", "open" = "open","high" = "high","low" = "low", "close"="close","volume" = "volume", "percent_change_price"= "PCP","percent_change_volume_over_last_wk"="PCVLW","previous_weeks_volume"="PWV","next_weeks_open"="NWO","next_weeks_close"="NWC","percent_change_next_weeks_price"="PCPNW","days_to_next_dividend"="DD","percent_return_next_dividend" = "PRND")

Removing '$' Sign : dow_jones_data <- dow_jones_data %>%mutate(NWC = as.numeric(gsub("\\$", "", NWC))),dow_jones_data <- dow_jones_data %>%mutate(NWO = as.numeric(gsub("\\$", "",

NWO))),dow_jones_data <- dow_jones_data %>% mutate(open = as.numeric(gsub("\\$", "",
open))),dow_jones_data <- dow_jones_data %>% mutate(high = as.numeric(gsub("\\$", "",
high))),dow_jones_data <- dow_jones_data %>% mutate(low = as.numeric(gsub("\\$", "",
low))),dow_jones_data <- dow_jones_data %>%mutate(close = as.numeric(gsub("\\$", "", close)))
dow_jones_data[,-(1:3)] **Normalize data** : normalized_data<-scale(dow_jones)
KMO Test :KMO(r=normalized_data) **Bartletts Test:** cortest.bartlett(normalized_data)
Correlation Matrix : corr_norm<-cor(normalized_data**) CorrPlot :** ggcorrplot::ggcorrplot(corr_norm)
Scree Plot: scree(normalized_data) Eigen Values : eigen_values_norm<-eigen (corr_norm)$values
var_explained_norm=(eign_values_norm/sum(eign_values_norm))
PCA FA: pca_norm<-fa(corr_norm,nfactors = 4,rotate = "varimax",n.obs = 720,cor=TRUE,fm="pa",max.iter =
1000,scores = "regression") **Communality:** pc_com_norm <-as.data.frame(unclass(pca_norm$communality))
pc_com_norm
MLE FA : ml_norm <- fa(corr_norm, nfactors = 5, rotate = "varimax", n.obs = 720 , corr = TRUE, fm = 'ml')
Communality : ml_Norm_com<-as.data.frame(ml_norm$communality)
Factor Diagram : fa.diagram(pca_norm)
Confirmatory FA
variables <-normalized_data[,-12]
model <- '
Factor1 =~ open+high+low+close+NWO+NWC
Factor2 =~ volume+PWV+PRND
Factor3 =~ PCVLW
Factor4 =~ PCPNW
' fit1<-sem(model,data = variables) summary(fit1,fit.measures=TRUE,standardized=TRUE)
semPaths(fit1, what = "col", whatLabels = "std", style = "mx", rotation = 1,
layout = "tree", nCharNodes = 8, shapeMan = "rectangle", sizeMan = 9, sizeMan2 = 5,edge.color="blue")