Mapping Systematic Risk and Industry Linkages with PCA and Factor Models

Evidence from U.S. Equities, 2005 – 2019

Zuoming Hu

Data Description

This project uses high-dimensional data analysis to study systematic risk and industry linkages in stock returns for portfolio optimization. The dataset includes monthly returns of 82 NYSE-listed companies and the S&P 500 index from January 2005 to December 2019. Time variables in the format "Y2005M01" were converted to the standard date format "2005-01-01" to align both datasets temporally. Market returns were transformed from percentages to decimals to match stock return formats, ensuring period-by-period alignment. Missing value analysis revealed no stock had more than 10% missing data, indicating good overall completeness. Remaining missing values were imputed using the mean to preserve time series continuity, ensuring data integrity for subsequent PCA and factor analysis.

Assumptions

This analysis rests on three assumptions: (1) the data are reliable, free of systematic measurement error, and accurately reflect market dynamics; (2) stock returns are driven by a limited number of underlying factors, making them suitable for principal component analysis and factor modeling; and (3) stocks within the same industry contribute equally to factor exposure, enabling cross-industry comparisons of systematic risk using average industry loadings. Based on these, we apply PCA and factor models to test whether the first principal component captures market trends, identify industries with highest systematic risk, and recommend five stocks for investment consideration.

Preliminary Analysis

Before conducting PCA, the characteristics of the original variables themselves should be explored first. The box plots of each variable are provided below.

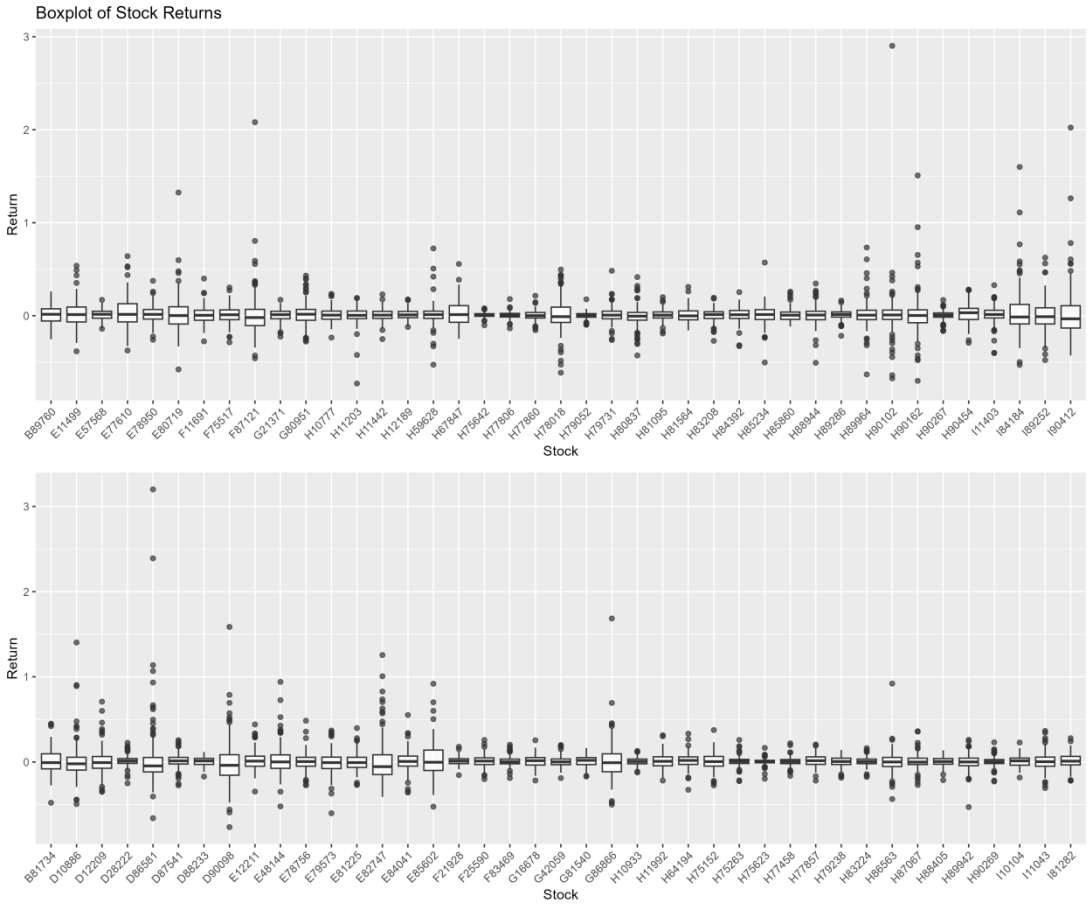


Figure 1: Box plots of Stock Returns

Central Tendency and Volatility Heterogeneity:
The plot highlights large differences across stocks.
• Median Returns: Medians reveal significant discrepancies. Some stocks have positive median values (boxes above zero), while others are near or below zero, reflecting differences in baseline performance across industries or business models.
• Volatility Differences: Box heights (IQR, 25th–75th percentiles) reveal clear differences. Narrow boxes indicate stable returns in defensive sectors like Transport & Utilities (E) and Services (I). Wider boxes show higher volatility in cyclical sectors such as Mining (B) and Manufacturing (D), where fluctuations are driven by commodity prices or the economic cycle.

Outlier Identification:
The data shows that there are extreme outliers for some certain stocks which are firm specific either very high or very low.
- Specific events such as mergers, surprises in earnings, actions of the firm due to the regulatory control,
- Certain specific shocks in the market like the one which happened in the year 2008, or the flash one of 2010
- These data anomalies are not impossible given known events. They result from real occurrences, and the documents preserve accurate risk profiles as a result.

Principal Components Analysis

Principal component analysis (PCA) was conducted on stock return data to identify the common sources of variation across both individual securities and industries.

Table 1: Summary of the principal component analysis of the stock

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	4.56416	0.25404	0.25404
PC2	2.24746	0.06160	0.31564
PC3	1.84943	0.04171	0.35735
PC4	1.67904	0.03438	0.39173
		……	
PC20	1.04527	0.01332	0.69703
PC21	1.02519	0.01282	0.70984
PC22	1.01173	0.01248	0.72233
PC23	0.98092	0.01173	0.73406
		……	

At the stock level, Table 1 shows that the first principal component (PC1) explains 25.4% of the total variance—substantially more than subsequent components. The second principal component (PC2) accounts for 6.16%, resulting in a cumulative explained variance of 31.56%. While 22 components have standard deviations greater than one, using the Kaiser criterion (eigenvalue > 1) would retain too many factors and is therefore inappropriate. Instead, the scree plot in Figure 3 shows an elbow at the second component, supporting the retention of PC1 and PC2.

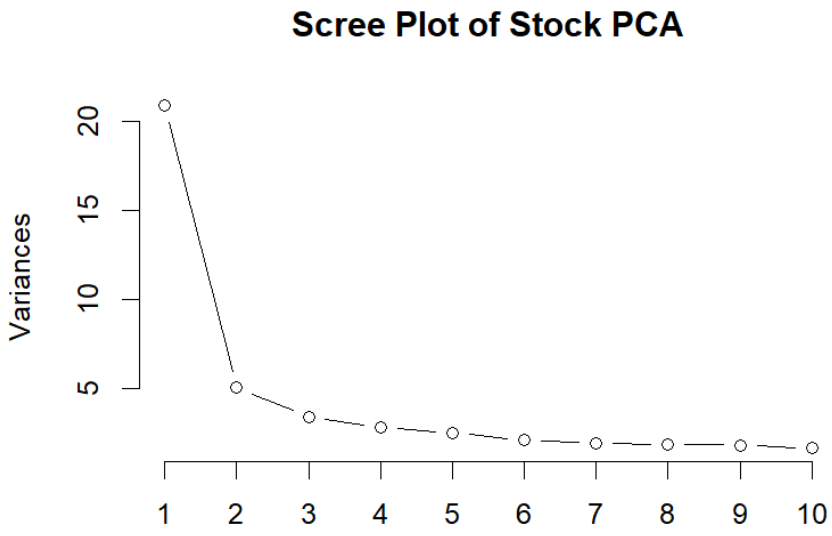


Figure 3: Scree Plot of Stock PCA

A correlation test between PC1 and market returns yielded a coefficient of –0.943, indicating a very strong inverse relationship. Given that principal component orientation is arbitrary, this reflects a strong positive association. Thus, PC1 at the stock level represents the market factor and captures systematic risk. Based on PC1 loadings, the five stocks most sensitive to the market are H89286, H88405, H90267, H83208, and D87541.

At the industry level, PCA reveals additional structure. Analyzing industry-average returns, the first principal component again dominates, reflecting common variation across sectors. This industry-level PC1 can thus be interpreted as a common factor driving systematic industry movements. In contrast, PC2 captures cross-industry divergence, highlighting sectors that deviate from the overall trend.

Table 2: The PCA loading of the industry

	PC1	PC2
Finance, Insurance & Real Estate	0.44503	0.09551
Manufacturing	0.36023	0.31877
Mining	0.27011	-0.83968
Retail Trade	0.38532	0.37641
Services	0.37065	-0.13900
Transportation and Public Utilities	0.43136	-0.12622
Wholesale Trade	0.35679	0.08522

As shown in Table 2, Mining, Services, and Transportation & Public Utilities have negative loadings on PC2, while Finance, Manufacturing, and Retail have positive loadings. This indicates that the former move inversely to the second factor, while the latter move in the same direction. Finance, Insurance & Real Estate shows the strongest exposure to PC1, and Mining has the highest loading on PC2, highlighting its distinct structural role.

Factor Analysis and Biplot

1. in factor model we selected two factors. We established this number based on the result that we got from scree plot when we were doing the PCA analysis. We saw a sharp decline at factor 2.
2.we calculated the covariance between factor 1 and the data from SampleF, which represented the market trending, to determine whether factor1 and market trending is related. And the result shows 0.8319093, which indicates the most loaded factor coincide with market movements. The result we got is over 0.8, which means our data has strong support to this conclusion.
3.from the biplot we analyzed which industry group has the greatest degree of loading on the first factor/second factor.

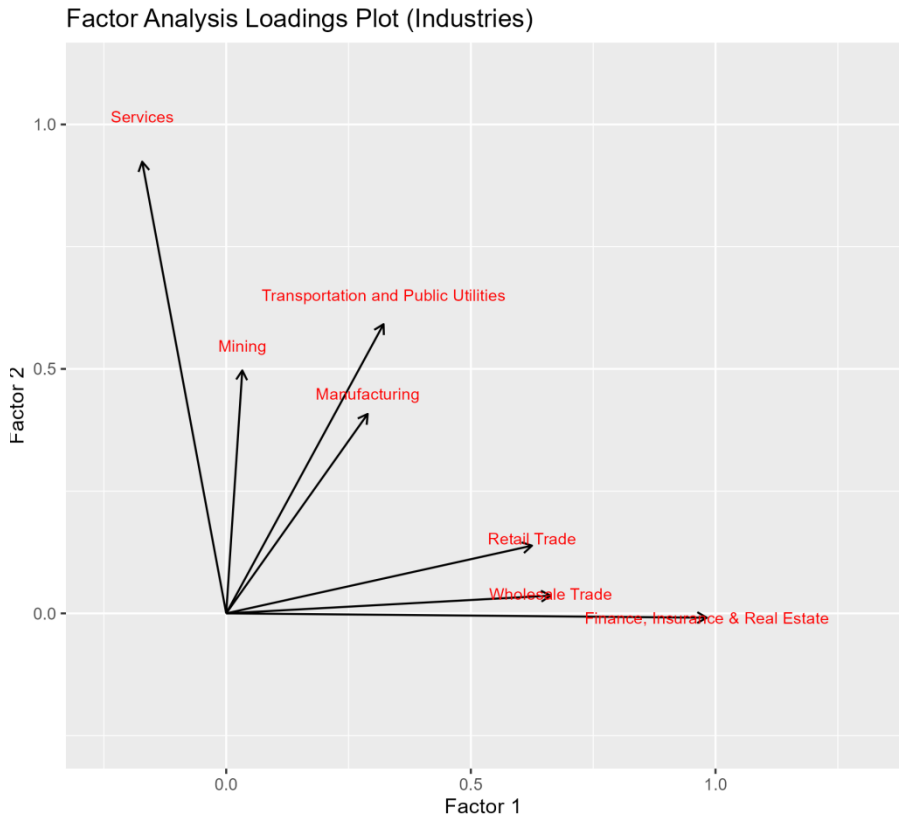


Figure 4: Factor Analysis Biplot

From the biplot we can see H industry group has the greatest degree on first factor, when switch to factor2, industry I seems to have the greatest degree on second factor. On top of this, we also acquired this result from directly calculating the loading.

Table 3:Industry Uniqueness and Factor Loadings

	uniqueness	Factor1 loading	Factor2 loading
H	0.0500000	0.98203363	-0.008869112
F	0.5204185	0.66307108	0.036258760
G	0.4512419	0.62493739	0.138535871
E	0.2412366	0.32154794	0.591554870
D	0.5606214	0.28913593	0.407913381
B	0.7257372	0.03287277	0.496940707
I	0.3717761	-0.17177580	0.923876252

4. Both from the table 3 and plot 4 we can conclude that industry I tend to move in the opposite directions of the factor1, industry H tend to move in the opposite directions of the factor2, but the tendency is subtle. H,F,G tend to have strong positive relation to Factor1. E,I,B tend to have strong positive relation to Factor2.
5. the industry with smaller uniqueness is vulnerable to market trending, which means to have larger systematic risk. Its industry H.
6. We selected the top five stocks with the smallest loadings.

Table 4: The top 5 stocks with the smallest loadings

Top Five smallest Stocks	uniqueness
H89286	0.06886122
H90267	0.96620514
H88405	0.26294437
H83208	0.31707714
H90269	0.39036131

Limitation

This analysis has several limitations. First, the two factors retained from the scree plot explain only about 35% of the total variance; capturing 70% requires over 20 factors, indicating that much of the data's information remains unaccounted for. Second, we assume equal contributions from all stocks within an industry. While large firms typically have greater influence, using a weighted average would provide a more realistic measure. Third, although no missing values were found, box plots reveal extreme returns for some stocks. These reflect real shocks such as the 2008 financial crisis and were retained to preserve the risk profile, but they may still distort PCA and factor loadings. Finally, PCA and factor models assume linear and stable relationships, yet market structures can evolve and involve nonlinear dynamics, meaning our static two-factor model may not fully capture systemic risks.

Summary

PCA results show that PC1 is highly correlated with market returns, confirming its representation of market factors and systematic risk. PC2 highlights cross-industry differences, revealing opposing industry dynamics and offering potential for diversification. At the industry level, PC1 captures broad common movements, while PC2 distinguishes risk exposures in sectors such as mining and finance.

These results yield three key insights. First, PC1 confirms the presence of strong market factors driving most stocks. Second, industry analysis shows that financial and utility sectors are more exposed to systemic risks, whereas industries with higher PC2 loadings exhibit divergent behavior and can hedge against market shocks. Third, by ranking stocks based on factor loadings, we identify five candidates most sensitive to the market, providing clear guidance for portfolio construction and risk management.

Appendix

```r
#check for missing values
colSums(is.na(SampleF))

to_date <- function(x) as.Date(sub("^Y(\\d{4})M(\\d+)$", "\\1-\\2-01", x))

#convert date column
Market$Date<-to_date(Market$Date)
SampleF$Date<-to_date(SampleF$Date)

#convert market yield
Market$MarketReturn <- Market$MarketReturn/100

#check the time is same
identical(SampleF$Date, Market$Date)

SampleF <- SampleF |>
  mutate(MarketReturn = (Market$MarketReturn)[match(Date, Market$Date)])
```

Preliminary Analysis

```r
#boxplot
numeric_df <- SampleF[ , -1]

long_df <- numeric_df |>
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value")

stock_vars <- unique(long_df$Variable)

half <- length(stock_vars) / 2
vars1 <- stock_vars[1:half]
vars2 <- stock_vars[(half+1):length(stock_vars)]

bp1 <- ggplot(filter(long_df, Variable %in% vars1),
              aes(x = Variable, y = Value)) +
  geom_boxplot(fill = "white", alpha = 0.7) +
  theme_gray() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Boxplot of Stock Returns",
       x = "Stock", y = "Return")

bp2 <- ggplot(filter(long_df, Variable %in% vars2),
              aes(x = Variable, y = Value)) +
  geom_boxplot(fill = "white", alpha = 0.7) +
  theme_gray() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Stock", y = "Return")

combined <- bp1 / bp2

ggsave("boxplot.png", combined, width = 12, height = 10, dpi = 300)
```

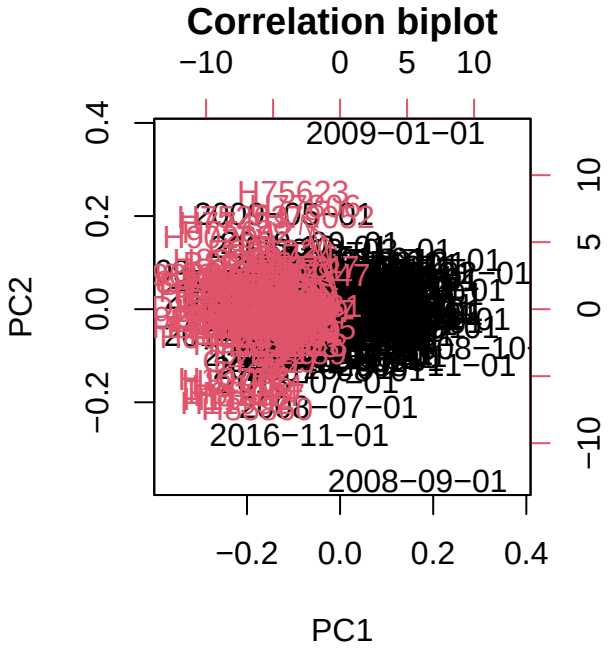
PCA

```
#simple exploratory analysis
STOCK <- "^[BCDEFGHI][0-9]{5}$"
stock_cols <- names(SampleF)[str_detect(names(SampleF), STOCK)]

#the rate of return of a stock in a certain period
long <-SampleF |>
  pivot_longer(matches(STOCK), names_to="stock", values_to="ret")

#the volatility of each stock was calculated and ranked
stats_stock <- long |>
  group_by(stock) |>
  summarise(mean   = mean  (ret,na.rm = T),
            median = median(ret,na.rm = T),
            sd     = sd    (ret,na.rm = T),
            mad    = mad   (ret,na.rm = T),
            .groups = "drop") |>
              arrange(desc(sd))

#analysis of volatility and return rate at the industry level
ind_names <- c(B = "Mining",
               C = "Construction",
               D = "Manufacturing",
               E = "Transportation and Public Utilities",
               F = "Wholesale Trade",
               G = "Retail Trade",
               H = "Finance, Insurance & Real Estate",
               I = "Services")

summ_industry <- stats_stock |>
  mutate(industry = substr(stock, 1, 1),
         industry_name = ind_names[industry]) |>
  group_by(industry, industry_name) |>
  summarise(mean_return = mean(mean, na.rm = TRUE),
            mean_vol    = mean(sd,   na.rm = TRUE),
            n_stocks    = n(),
            .groups = "drop") |>
              arrange(desc(mean_vol))

#the ranking of the correlation between industries and the overall market
cor_industry <- SampleF |>
  summarise(across(matches(STOCK), ~ cor(.x, MarketReturn,
                                         use = "complete.obs"))) |>
  pivot_longer(everything(), names_to="stock", values_to="corr_with_market") |>
  mutate(industry = substr(stock,1,1),
         industry_name = ind_names[industry]) |>
  group_by(industry, industry_name) |>
  summarise(mean_corr = mean(corr_with_market, na.rm=TRUE), .groups="drop") |>
              arrange(desc(mean_corr))

#a list of stocks that are most in sync with the market
cor_stock_market <- SampleF |>
  summarise(across(all_of(stock_cols), ~ cor(.x, MarketReturn,
                                             use = "complete.obs"))) |>
  pivot_longer(everything(), names_to = "stock", values_to = "rho") |>
                arrange(desc(rho))
```

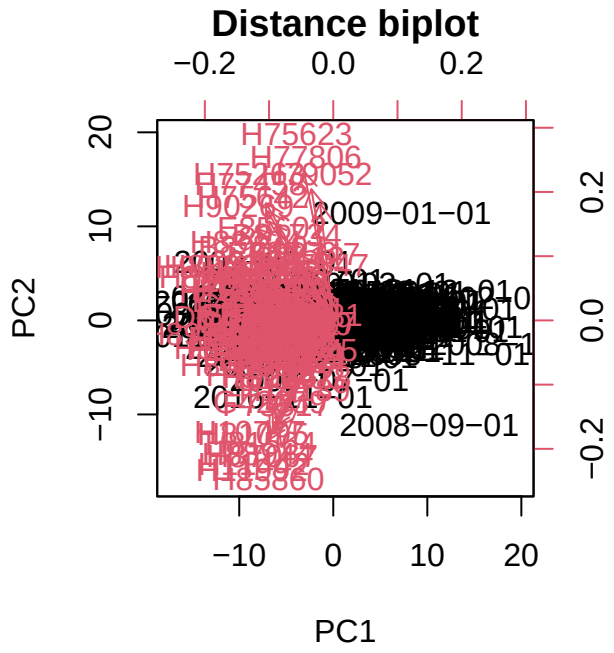
```r
#PCA
stocks_for_pca <- SampleF |>
  dplyr::select(Date,matches(STOCK)) |>
  distinct(Date, .keep_all = TRUE) |>
  column_to_rownames("Date")

stocks_pca <- prcomp(stocks_for_pca, scale. = TRUE)
summary(stocks_pca)
pca_summary <- summary(stocks_pca)

#export data
importance <- as.data.frame(pca_summary$importance)
head(importance)
write.csv(importance, "pca_summary.csv", row.names = TRUE)
```

```r
#screeplot
screeplot(stocks_pca, type = 'lines')
```

```r
#biplot
biplot(stocks_pca, main = "Correlation biplot")
```



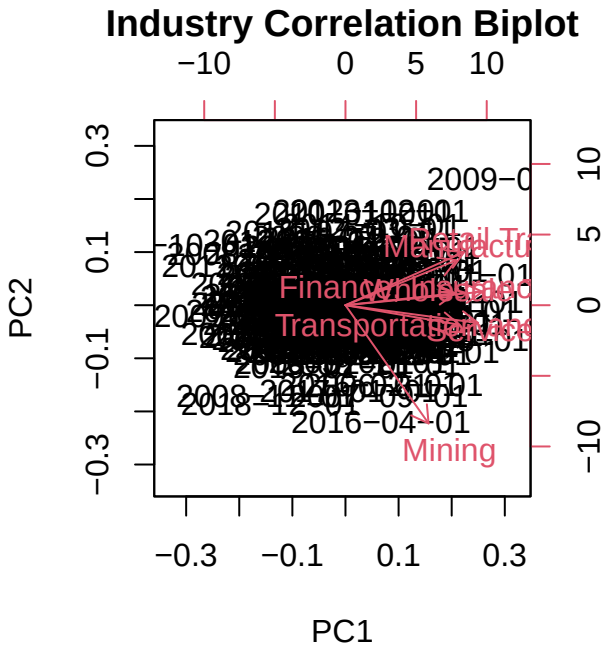

```r
biplot(stocks_pca, scale = 0, main = "Distance biplot")
```



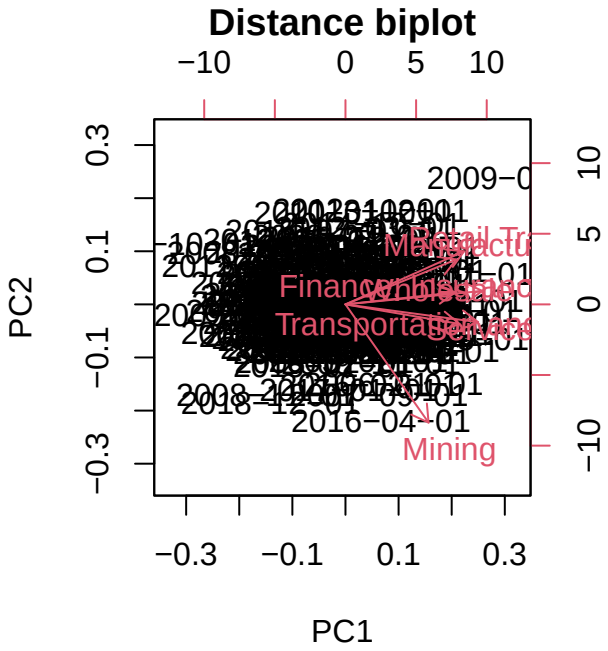

```
#extract the PC1 score and compare it with the market
pc1_scores <- stocks_pca$x[, 1]
cor_pc1_mkt <- cor(pc1_scores, SampleF$MarketReturn, use = "complete.obs")

#the 5 stocks with the highest correlation to PC1
pc1_corr <- apply(stocks_for_pca,
                  2,
                  function(x) cor(x, pc1_scores,
                                  use="complete.obs"))
sort(pc1_corr)[1:5]
```

```
##     H89286     H90267     H88405     H83208     D87541
## -0.9562255 -0.8671894 -0.8423259 -0.8055117 -0.7405279
```

```
#PCA by industry
industry_returns <- SampleF |>
  select(Date, all_of(stock_cols)) |>
  pivot_longer(-Date, names_to="stock", values_to="ret") |>
  mutate(industry = substr(stock, 1, 1),
         industry_name = ind_names[industry]) |>
  group_by(Date, industry_name) |>
  summarise(ind_ret = mean(ret, na.rm=TRUE), .groups="drop") |>
  pivot_wider(names_from=industry_name, values_from=ind_ret) |>
  column_to_rownames("Date")

industry_pca <- prcomp(industry_returns, scale. = TRUE)

#biplot
biplot(industry_pca, main="Industry Correlation Biplot")
```




```
biplot(industry_pca, scale = 1, main = "Distance biplot")
```

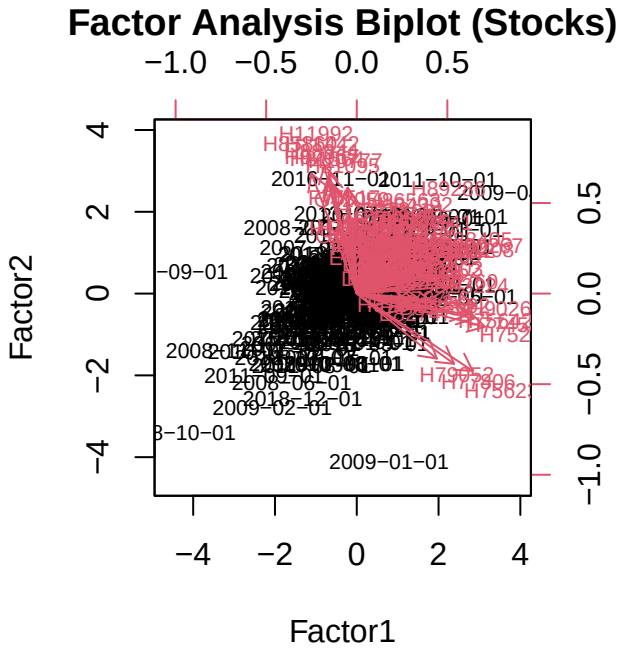



```
#the PCA loading of the industry
industry_loadings <- industry_pca$rotation
industry_loadings <- industry_loadings[, 1:2]
industry_loadings_df <- as.data.frame(industry_loadings)
industry_loadings_df <- tibble::rownames_to_column(industry_loadings_df,
                                                   var = "Industry")

#export data
write.csv(industry_loadings_df, "industry_pca_loadings.csv", row.names = FALSE)
```

Factor Analysis

```r
#prepare data (individual stock)
Factor <- stocks_for_pca

facto <- factanal(Factor,
                  factors = 2,
                  scores = 'Bartlett',
                  rotation = 'promax',
                  lower = 0.05)
```

```r
#individual stock biplot
scores <- facto$scores[, 1:2, drop = FALSE]
loadings <- facto$loadings[, 1:2]
biplot(scores, loadings, cex = 0.7, main = "Factor Analysis Biplot (Stocks)")
```




```r
#correlation with Market Returns
fa1_scores <- scores[,1]
cor_fa1_mkt <- cor(fa1_scores,
                   SampleF$MarketReturn[match(rownames(scores),
                                              SampleF$Date)],
                   use = "complete.obs")

#individual stock uniqueness and industry average
uni <- facto$uniquenesses
stock_vec <- names(uni)
industry_code <- substr(stock_vec, 1, 1)
industry_name <- ind_names[industry_code]

industry_uniqueness <- aggregate(uni ~ industry_name,
                                 data = data.frame(stock = stock_vec,
                                                   industry_name = industry_name,
                                                   uni = uni),
                                 FUN = mean)
industry_uniqueness <- industry_uniqueness[order(industry_uniqueness$uni), ]
industry_uniqueness
```

```
##                          industry_name       uni
## 1    Finance, Insurance & Real Estate 0.6204843
```

```
## 3                              Mining 0.7286184
## 2                       Manufacturing 0.7678617
## 4                        Retail Trade 0.7702905
## 7                     Wholesale Trade 0.7738501
## 6 Transportation and Public Utilities 0.7823407
## 5                            Services 0.7934785
```

```r
#measure market sensitivity based on factor loading
load_df <- as.data.frame(unclass(loadings))
load_df$stock <- rownames(load_df)
load_df$industry <- ind_names[substr(load_df$stock, 1, 1)]
load_df <- load_df[, c("stock", "industry",
                       colnames(load_df)[1:(ncol(load_df)-2)])]


#top 5 stocks with the highest and lowest exposure to Factor 1
top5_f1 <- load_df[order(-abs(load_df$Factor1)), ][1:5, ]

bottom5_f1 <- load_df[order(abs(load_df$Factor1), -abs(load_df$Factor2)), ][1:5, ]

print(top5_f1)
```

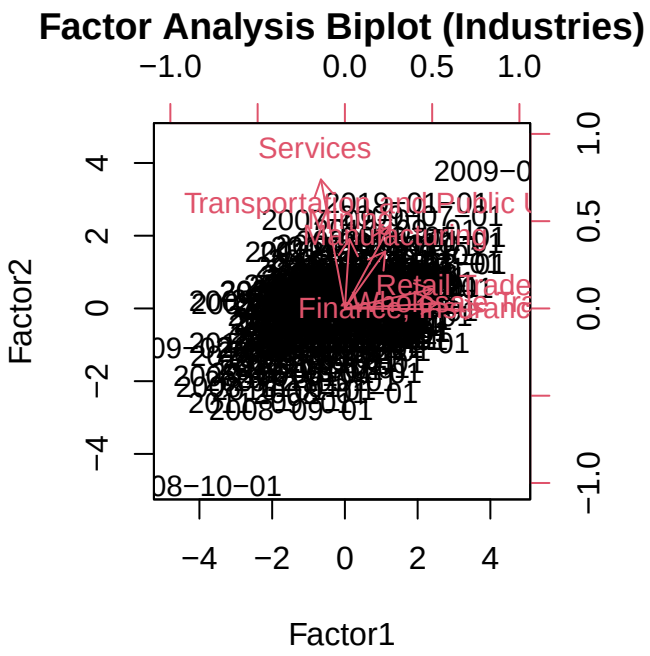
```
##         stock                          industry   Factor1     Factor2
## H75263 H75263 Finance, Insurance & Real Estate 0.8843049 -0.22844525
## H90269 H90269 Finance, Insurance & Real Estate 0.8241392 -0.08094086
## H77458 H77458 Finance, Insurance & Real Estate 0.8090873 -0.17158606
## H75623 H75623 Finance, Insurance & Real Estate 0.7997706 -0.53340066
## H75642 H75642 Finance, Insurance & Real Estate 0.7667308 -0.14693203
```

```r
print(bottom5_f1)
```

```
##         stock                          industry     Factor1   Factor2
## H11203 H11203 Finance, Insurance & Real Estate -0.01350967 0.3697388
## G16678 G16678                     Retail Trade -0.02044450 0.3226840
## G42059 G42059                     Retail Trade -0.02774704 0.5188800
## H89942 H89942 Finance, Insurance & Real Estate  0.03500319 0.3961619
## H64194 H64194 Finance, Insurance & Real Estate  0.04757747 0.3014970
```

```r
#conduct factor analysis on the average industry returns
FactorI <- industry_returns
factoI <- factanal(FactorI,
                   factors = 2,
                   scores = 'Bartlett',
                   rotation = 'promax',
                   lower = 0.05)
```

```r
#industry factor loadings plot
scoresI <- factoI$scores[, 1:2, drop = FALSE]
loadingsI <- factoI$loadings[, 1:2]

biplot(scoresI, loadingsI, cex = 0.9, main = "Factor Analysis Biplot (Industries)")
```




```
loadingsI <- as.data.frame(unclass(factoI$loadings[, 1:2]))
colnames(loadingsI) <- c("Factor1","Factor2")
loadingsI$Industry <- rownames(loadingsI)

#ggplot
fl <- ggplot(loadingsI, aes(x = 0, y = 0,
                      xend = Factor1, yend = Factor2)) +
  geom_segment(arrow = arrow(length = unit(0.2, "cm")),
               colour = "black", linewidth = 0.5) +
  geom_text(aes(x = Factor1 * 1, y = Factor2 * 1.1, label = Industry),
            color = "red", size = 3) +
  coord_equal(xlim = c(-0.25, 1.3), ylim = c(-0.25, 1.1)) +
    scale_x_continuous(breaks = seq(0, 1.0, by = 0.5)) +
  scale_y_continuous(breaks = seq(0, 1.0, by = 0.5)) +
  labs(title = "Factor Analysis Loadings Plot (Industries)",
       x = "Factor 1", y = "Factor 2") +
  theme_gray()

ggsave("factor_loadings.png", plot = fl, width = 8, height = 6, dpi = 300)
```

```
#which industries' fluctuations are mainly explained by the common factor
#identify the industries with the most significant loading on factor 1 and factor 2
uniI <- factoI$uniquenesses
uniI <- sort(uniI)

loadingsI <- factoI$loadings[, 1:2]

loadsI <- as.data.frame(unclass(loadingsI))
colnames(loadsI) <- c("Factor1","Factor2")
loadsI$industry <- rownames(loadsI)

top_f1_ind <- loadsI[order(-abs(loadsI$Factor1)), ][1:3, ]
print(top_f1_ind)
```

```
##                                 Factor1      Factor2
## Finance, Insurance & Real Estate 0.9820336 -0.008869112
## Wholesale Trade                  0.6630711  0.036258760
```

```
## Retail Trade                     0.6249374  0.138535871
##                                                                industry
## Finance, Insurance & Real Estate Finance, Insurance & Real Estate
## Wholesale Trade                                        Wholesale Trade
## Retail Trade                                              Retail Trade
```

```
top_f2_ind <- loadsI[order(-abs(loadsI$Factor2)), ][1:3, ]
print(top_f2_ind)
```

```
##                                      Factor1   Factor2
## Services                          -0.17177580 0.9238763
## Transportation and Public Utilities  0.32154794 0.5915549
## Mining                                0.03287277 0.4969407
##                                                                industry
## Services                                                       Services
## Transportation and Public Utilities Transportation and Public Utilities
## Mining                                                           Mining
```

```
#the top 5 stocks with the smallest loadings
stk_uni <- sort(facto$uniquenesses)
top5_u <- head(stk_uni, 5)
tab4 <- data.frame(Stock = names(top5_u), Uniqueness = as.numeric(top5_u))
```