homework 7

Shihan Ban, Yi Cao, Shri Lekkala, Ningxin Zhang

2024-05-10

Details

What are the latent factors of international currency pricing? And how do these factor move against US equities?

We're going to investigate underlying factors in currency exchange rates and regress the S&P 500 onto this information.

FX data is in FXmonthly.csv. SP500 returns are in sp500csv. Currency codes are in currency codes.txt.

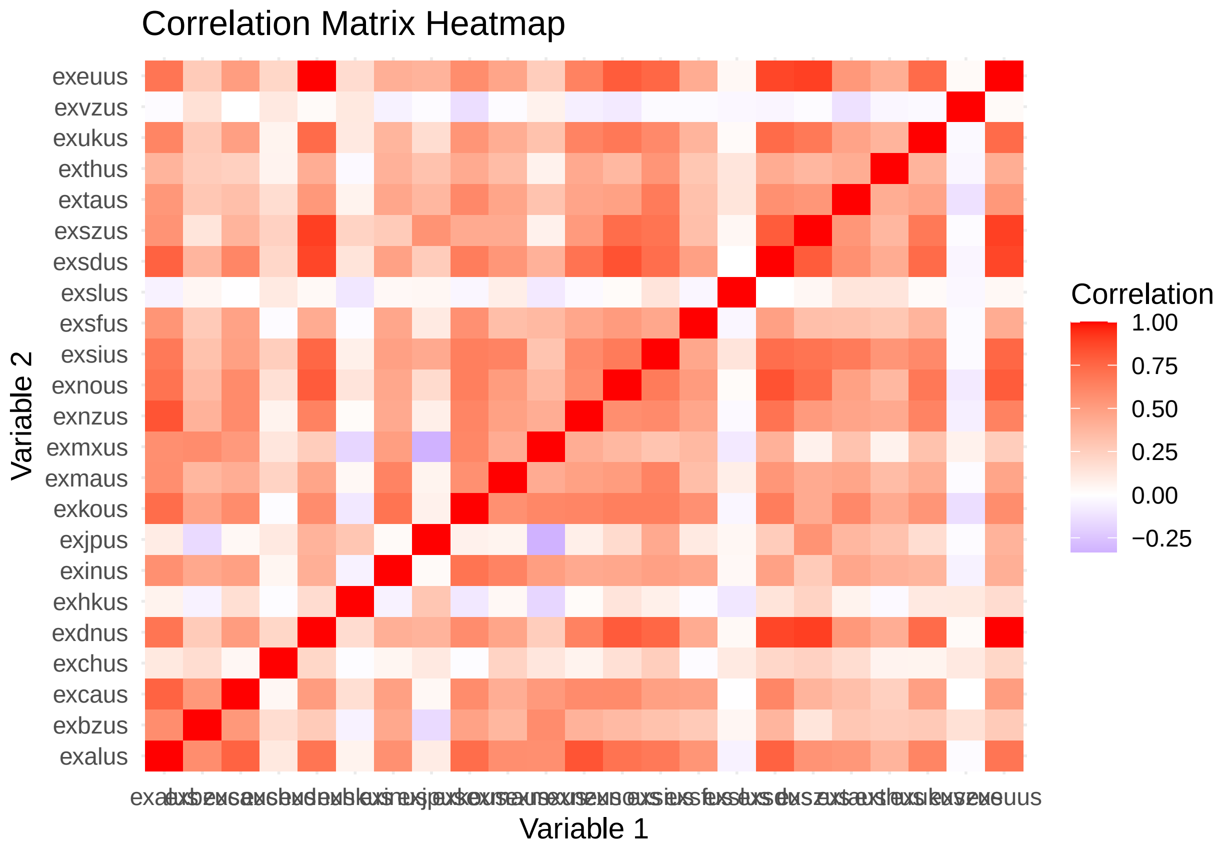
```
library(textir)
library(dplyr)
library(tidyr)
library(knitr)
library(kableExtra)
library(ggplot2)
library(reshape2)
fx <- read.csv("C:/Users/23713/Downloads/FXmonthly.csv")
fx <- (fx[2:120,]-fx[1:119,])/(fx[1:119,])
```

question 1

Discuss correlation amongst dimensions of fx. How does this relate to the applicability of factor modelling?

```
cor_matrix <- cor(fx, use = "complete.obs")

melted_cor_matrix <- melt(cor_matrix)
ggplot(melted_cor_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  labs(title = "Correlation Matrix Heatmap", x = "Variable 1", y = "Variable 2", fill = "Correlation")
```



The color gradient of the heatmap spans from red to white to blue, signifying correlations from 1.00 to -0.25. Red signifies a strong positive correlation, white denotes no correlation, and blue indicates negative correlations. The prevalence of red hues across the matrix suggests numerous variables are positively correlated to a moderate or strong extent. Darker red clusters hint at variables strongly correlated, possibly indicating their joint movement or shared influences. Sparse blue areas suggest minimal negative correlations between variables, implying few pairs show inverse relationships in the dataset. As expected, the diagonal, reflecting self-correlations, shines bright red, indicating a perfect correlation of 1.00.

applicability of factor modelling

The prevalence of high positive correlations, indicated by the dark red areas, implies shared underlying factors among these variables, making it conducive to factor modeling. Factor models aim to diminish dimensionality by identifying a handful of underlying factors that elucidate the observed correlations across multiple variables. Given the strong correlations among many variables, principal component analysis (PCA) or another factor analysis method could efficiently reduce dataset dimensions by extracting key components that encapsulate the majority of variability.

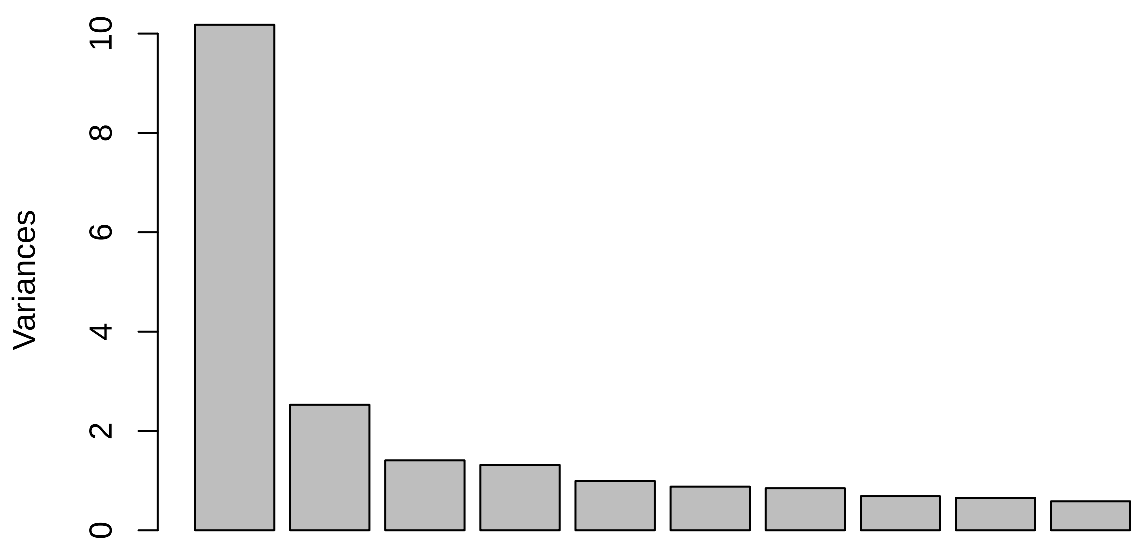
question 2

Fit, plot, and interpret principal components.

```r
pcafx<- prcomp(fx, scale=TRUE)
summary(pcafx)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.1904 1.5905 1.18680 1.14792 0.99740 0.93815 0.92009
## Proportion of Variance 0.4425 0.1100 0.06124 0.05729 0.04325 0.03827 0.03681
## Cumulative Proportion  0.4425 0.5525 0.61377 0.67107 0.71432 0.75258 0.78939
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.82835 0.80841 0.76390 0.69185 0.65917 0.58024 0.56012
## Proportion of Variance 0.02983 0.02841 0.02537 0.02081 0.01889 0.01464 0.01364
## Cumulative Proportion  0.81923 0.84764 0.87301 0.89382 0.91271 0.92735 0.94099
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.55254 0.50190 0.44624 0.41834 0.38808 0.33724 0.30771
## Proportion of Variance 0.01327 0.01095 0.00866 0.00761 0.00655 0.00494 0.00412
## Cumulative Proportion  0.95427 0.96522 0.97388 0.98149 0.98803 0.99298 0.99709
##                          PC22    PC23
## Standard deviation     0.2580 0.01557
## Proportion of Variance 0.0029 0.00001
## Cumulative Proportion  1.0000 1.00000
```

```r
plot(pcafx, main="")
```



The analysis of variance explained by principal components (PCs) reveals that PC1 accounts for a significant portion of dataset variance, specifically 44.25%, indicating its ability to capture a substantial part of the data's variability. PC2 contributes an additional 11.00%, resulting in a cumulative variance explained of 55.25% for the first two components. Subsequent components show diminishing contributions to variance, with cumulative proportions gradually increasing towards 100% by PC21. By PC5, over 75% of the dataset's variance is explained, indicating that the first five components capture the majority of information, while PC8 reaches approximately 82%, suggesting that additional components beyond this point contribute incrementally less to overall data variability. Visually, the scree plot underscores this trend, displaying a steep drop after the first component and another noticeable decrease after the second, aligning with the 'elbow' method's indication that crucial information is primarily captured within the initial components, particularly by the fifth component.

question 3

Regress SP500 returns onto currency movement factors, using both 'glm on first K' and lasso techniques.

Use the results to add to your factor interpretation.

```
sp <- read.csv("C:/Users/23713/Downloads/sp500.csv")

# Find how many components sum to at least 90% of the variance
cumulative_variance <- summary(pcafx)$importance[3,]
k <- which(cumulative_variance >= 0.9)[1]

print(k)
```

```
## PC12
##   12
```

```
sp500 <- sp$sp500

fx1 <- predict(pcafx)
zdf <- as.data.frame(fx1)

kfits <- lapply(1:12,
    function(K) glm(sp500~., data=zdf[,1:K,drop=FALSE]))

aicc <- sapply(kfits, AICc) # apply AICc to each fit
which.min(aicc)
```

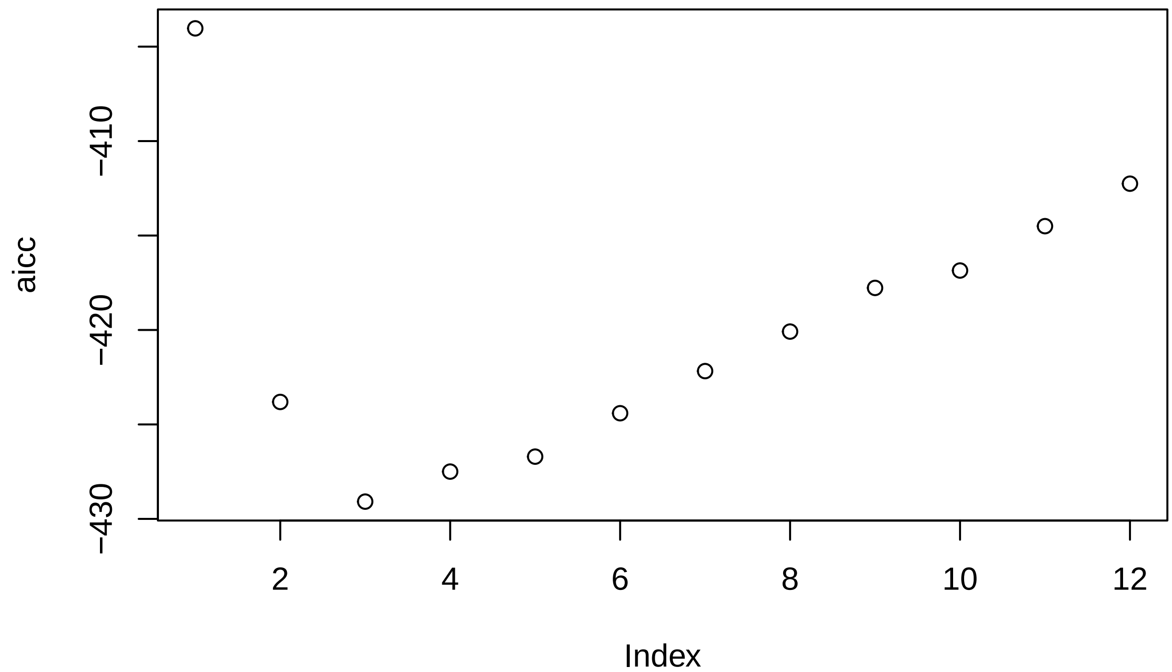
```
## [1] 3
```

```
bic <- sapply(kfits, BIC)

which.min(bic)
```

```
## [1] 3
```

```
plot(aicc)
```



compare to the output by the lasso

```
set.seed(142)
lassoPCR <- cv.gamlr(x=fx1, y=sp500, nfold=20)
coef(lassoPCR)
```

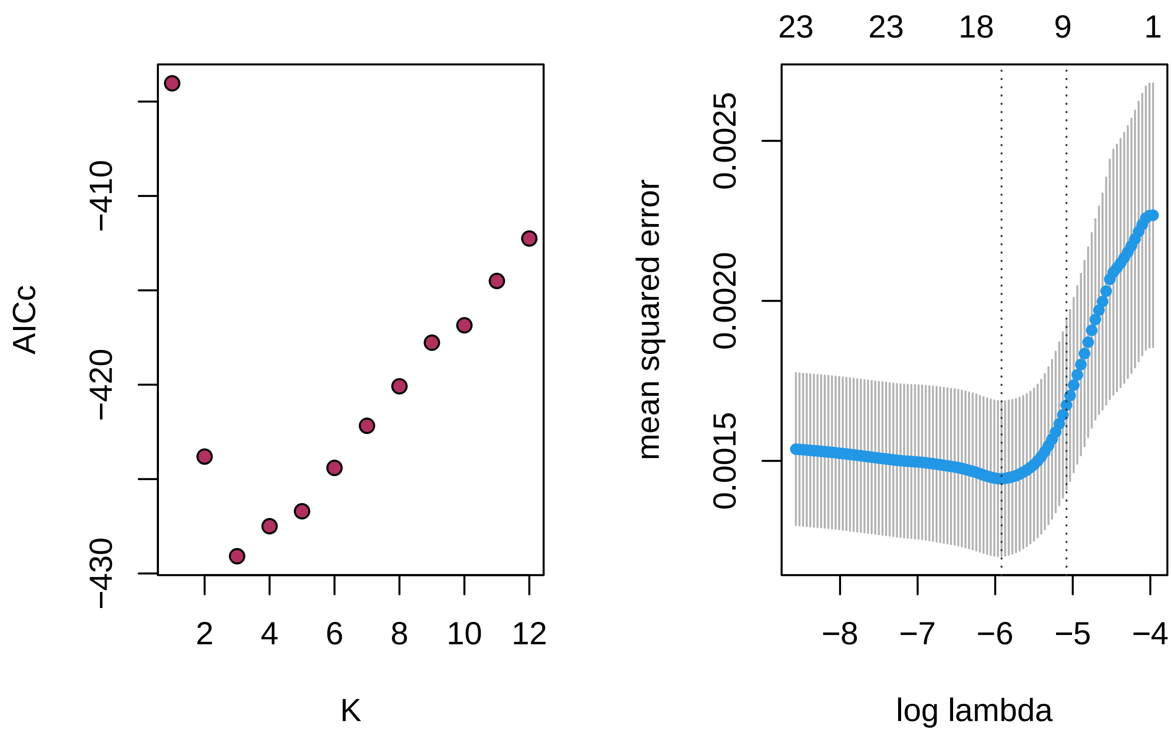
```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                    seg25
## intercept  0.0004430924
## PC1        0.0040178644
## PC2       -0.0072552637
## PC3       -0.0029511302
## PC4        .
## PC5        .
## PC6        .
## PC7        .
## PC8        .
## PC9        .
## PC10       .
## PC11       .
## PC12       .
## PC13       .
## PC14       .
```

```
## PC15       -0.0017412344
## PC16        .
## PC17       -0.0072242161
## PC18        .
## PC19        .
## PC20       -0.0093955611
## PC21       -0.0005361697
## PC22        .
## PC23        0.1621996215
```

```r
par(mfrow=c(1,2))

plot(aicc, pch=21, bg="maroon", xlab="K", ylab="AICc")

plot(lassoPCR)
```

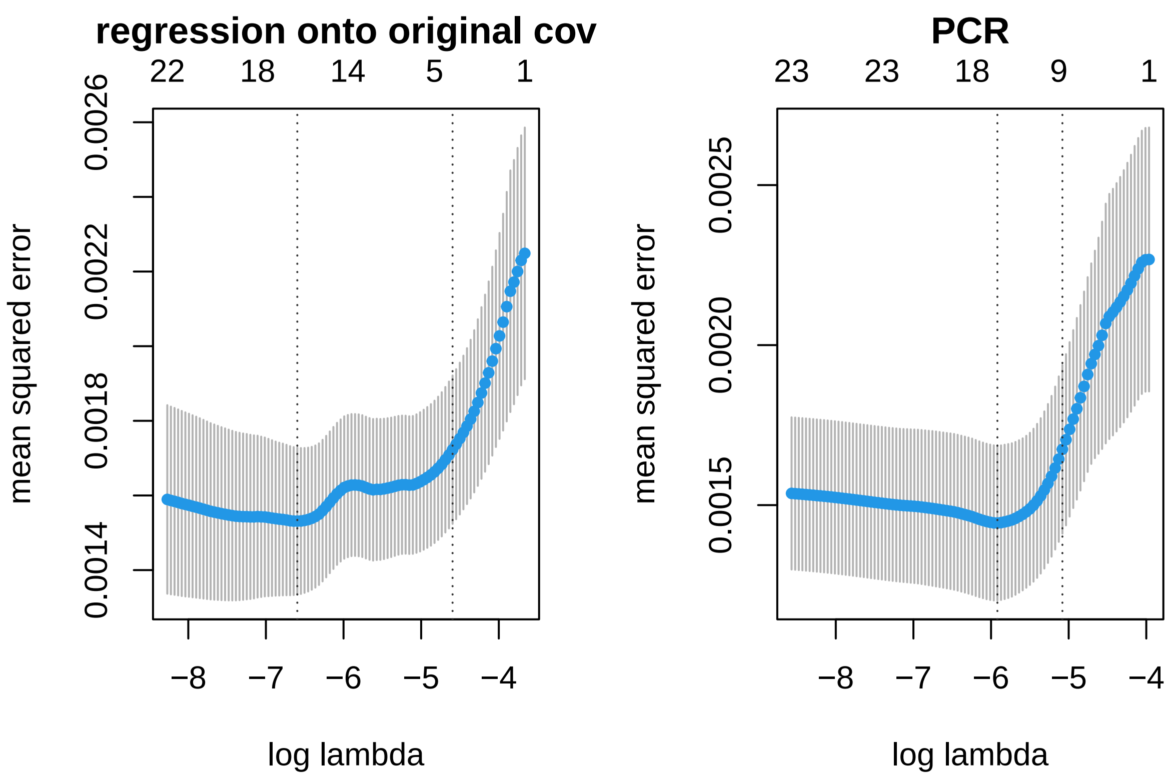


Combining the results from PCA and regression models, we gain a comprehensive understanding of the relationship between S&P 500 returns and currency movement factors. PCA reveals that PC1 accounts for a substantial portion of dataset variance (44.25%), with PC2 contributing an additional 11.00%, cumulatively explaining 55.25% of the variability. Subsequent components exhibit diminishing contributions, with over 75% of variance explained by PC5 and approximately 82% by PC8, as depicted in the scree plot. Model selection based on AICc and BIC suggests that a model with 3 principal components strikes the optimal balance between complexity and fit. Lasso regression identifies significant components like PC22, highlighting their strong influence on S&P 500 returns. These findings underscore the importance of specific currency movement factors, potentially representing key economic or financial indicators, in explaining fluctuations in S&P 500 returns, offering valuable insights for investment decisions and financial strategy development.

question 4

Fit lasso to the original covariates and describe how it differs from PCR here.

```
## compare to an un-factorized lasso
par(mfrow=c(1,2))
lasso <- cv.gamlr(x=as.matrix(fx), y=sp500, nfold=20)
plot(lasso, main="regression onto original cov")
plot(lassoPCR, main="PCR")
```



In the provided plots comparing Lasso regression on original covariates and Principal Component Regression (PCR), key differences emerge in model complexity and prediction accuracy. Lasso directly applies to the original variables, offering a sparse model that emphasizes variable selection, potentially enhancing interpretability by identifying key predictors. In contrast, PCR reduces dimensionality by focusing on principal components, which represent major variance directions, potentially reducing the impact of multicollinearity and noise. The plots illustrate that PCR might start with a lower mean squared error (MSE) at higher lambda values, suggesting better initial robustness against overfitting, whereas Lasso shows a continuous decrease in MSE with more complex models until stabilizing. The choice between these methods should consider the specific data characteristics and analytical goals: Lasso for direct variable impact analysis and PCR for capturing underlying data patterns when dealing with high-dimensional or multicollinear data. Overall, cross-validation results such as those shown guide the selection process by empirically demonstrating each method's predictive performance and complexity trade-offs.

to answer the beginning question

PCA reveals that the first principal component (PC1) explains a significant 44.25% of the dataset's variance, indicating its ability to capture a substantial portion of the data's variability, primarily reflecting major economic or financial indicators influencing currency values. This component, along with others contributing to a cumulative variance explanation of 90% by the 12th component, underscores the significant underlying structures in currency exchange rates. In terms of comparing with the trends of the U.S. stock market, regressing S&P 500 returns onto these currency movement factors using both generalized linear models (GLM) and Lasso techniques suggests that only a few principal components are critically relevant. Specifically, a model containing three principal components is identified as optimal based on AICc and BIC, indicating their substantial explanatory power over S&P 500 fluctuations. The Lasso results further emphasize this by identifying particular components like PC22 that significantly influence S&P 500 returns, pointing towards specific currency factors that might correlate with or predict U.S. equity market movements. These insights could be pivotal for financial strategy and investment decision-making, reflecting the intertwined dynamics of international currencies and the U.S. stock market.