BUS 41201 Homework 7 Assignment

Group 11

2024-05-12

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.

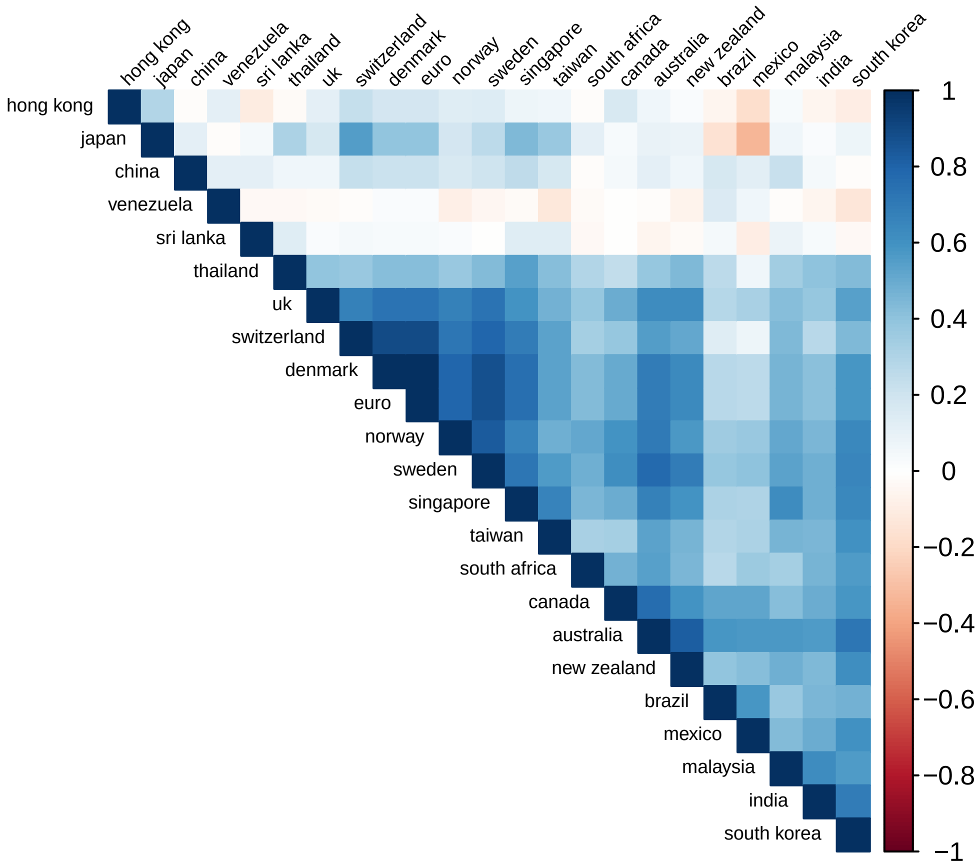
Translate the prices to 'returns' via

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

```
setwd("C:/Users/user/Desktop/Big Data/HW/week8")
fx <- read.csv("FXmonthly.csv")
fx <- (fx[2:120,]-fx[1:119,])/(fx[1:119,])
```

```
# Read the currency codes file
currency <- read.table("currency_codes.txt", header = FALSE, col.names = c("Code", "Country"))

# Extract the currency pair abbreviations from the column names
currency_pairs <- substr(colnames(fx), 3, 4)

# Match the currency pair abbreviations with the currency codes
matched_currencies <- match(currency_pairs, currency$Code)

# Add the corresponding country names to the FX dataset
fx_countries <- currency$Country[matched_currencies]

# Add the matched country names as column names to the FX dataset
colnames(fx) <- fx_countries
```

```
# Calculate the correlation matrix
correlation_matrix <- cor(fx)

# Plot the correlation matrix with larger cell size
corrplot(correlation_matrix, method = "color", type = "upper", order = "hclust", tl.col = "black", tl.s
```



Correlation among FX dimensions reflects the relationships between different currency pairs. Strong positive correlations suggest pairs tend to move together, indicating shared underlying factors. Factor modeling, like PCA, can effectively capture this shared variation.

[2] Fit, plot, and interpret principal components.

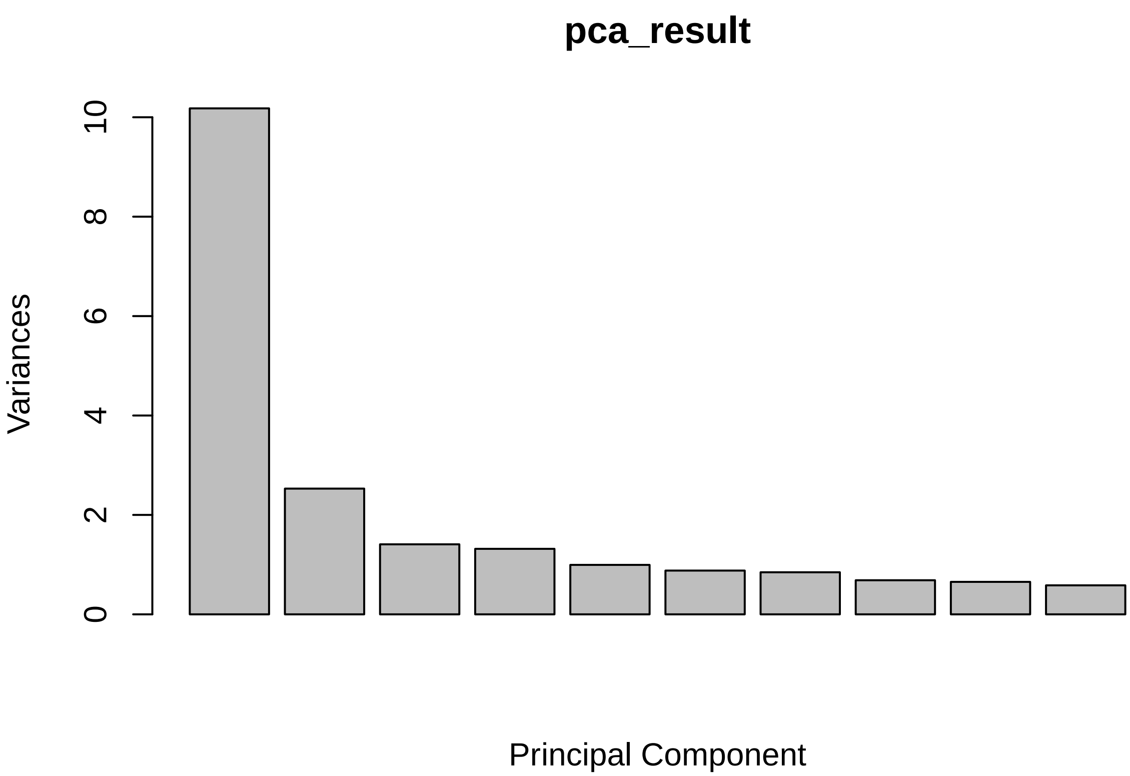
```
# Assuming 'fx' is your FX dataset
pca_result <- prcomp(fx, scale. = TRUE)

summary(pca_result)
```

```
## 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
```

- Proportion of Variance: This represents the proportion of total variance in the data explained by each principal component. PC1 explains 44.25% of the total variance, PC2 explains 11.00%, and so on. Higher proportions suggest greater importance of the corresponding component in capturing the variability of the data.

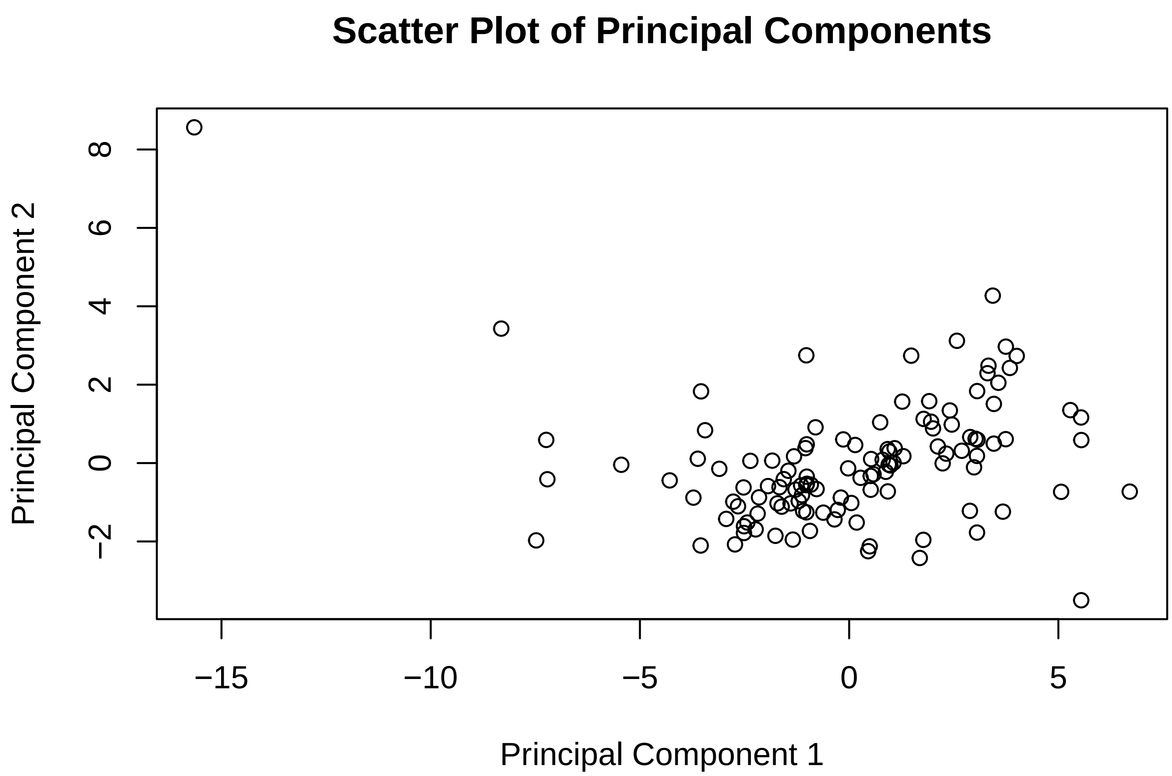
```r
# The scree plot
plot(pca_result, xlab = "Principal Component")
```



The scree plot helps in determining the number of principal components to retain in the analysis. In an ideal scenario, the scree plot displays a steep decline in the proportion of variance explained as we move from the first component to subsequent components. The "elbow" point, where the decline becomes less steep, often indicates the number of principal components to retain.

In this case, only the first bar (PC1) is high, it suggests that the first principal component explains a significant portion of the variance in the data compared to the other components.

```r
# Extract the scores of the observations along the principal components
scores <- pca_result$x

# Plot the scores of the observations on the first two principal components
plot(scores[,1], scores[,2],
     xlab = "Principal Component 1", ylab = "Principal Component 2",
     main = "Scatter Plot of Principal Components")
```



Most of the dots in the scatter plot are clustered in the bottom right corner, it suggests a strong negative correlation between the first two principal components.

[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.

```
sp500 <- read.csv("sp500.csv")

# Extract the first K principal components
K <- 3  # Choose the number of principal components to include
predict_pca <-predict(pca_result)
currency_factors <- as.data.frame(predict_pca[, 1:K])
```

```
# GLM on First K Technique
sp500 <- sp500$sp500
glm_sp <- glm(sp500 ~ ., data = currency_factors, family = gaussian)
summary(glm_sp)
```

```
##
## Call:
## glm(formula = sp500 ~ ., family = gaussian, data = currency_factors)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.10729  -0.02094   0.00126   0.02447   0.08594
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0004431  0.0035581   0.125  0.90111
## PC1          0.0059741  0.0011200   5.334 4.87e-07 ***
## PC2         -0.0111795  0.0022466  -4.976 2.30e-06 ***
## PC3          0.0082100  0.0030107   2.727  0.00739 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001506534)
##
##     Null deviance: 0.26463  on 118  degrees of freedom
## Residual deviance: 0.17325  on 115  degrees of freedom
## AIC: -429.62
##
## Number of Fisher Scoring iterations: 2
```

GLM on First K Technique:

- PC1, PC2, PC3 Coefficients: These coefficients represent the estimated effect of each principal component on the SP500 returns, holding other variables constant.
- PC1 coefficient (0.0059741): A one-unit increase in PC1 is associated with an increase in SP500 returns by 0.0059741 units, holding other factors constant.
- Significance: PC1 and PC2 are statistically significant at the 0.01 level, while PC3 is significant at the 0.05 level. This indicates that these principal components have a significant impact on SP500 returns.

```
# Lasso Technique
lasso_model <- cv.glmnet(x=predict_pca, y=sp500, nfold=20)

coef(lasso_model)
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  0.0004430924
## PC1          0.0038271320
## PC2         -0.0068726594
## PC3          0.0024383962
## PC4          .
## PC5          .
## PC6          .
## PC7          .
## PC8          .
## PC9          .
## PC10         .
## PC11         .
## PC12         .
## PC13         .
## PC14         .
## PC15        -0.0006399351
## PC16         .
## PC17         0.0058605697
## PC18         .
## PC19         .
```

```
## PC20         0.0075911504
## PC21         .
## PC22         .
## PC23        -0.1231162884
```

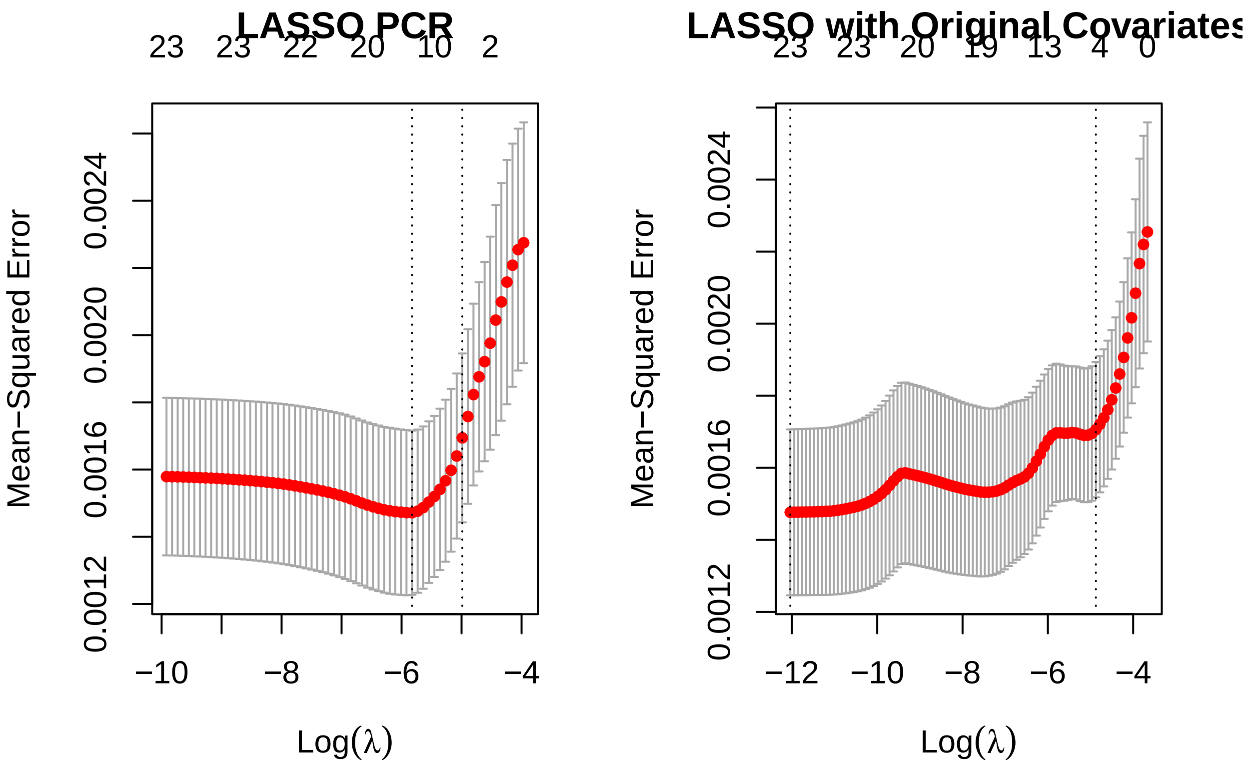
- The Lasso technique performs variable selection and regularization to prevent overfitting by shrinking coefficients towards zero.
- The output shows the estimated coefficients for each principal component chosen by the Lasso model.
- Some coefficients are set to zero, indicating that the corresponding principal components were not selected by the Lasso model. This implies that those components have negligible impact on predicting SP500 returns according to the Lasso model.

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

```
# Assuming 'fx' contains your original covariates
lasso_model_original <- cv.glmnet(x = as.matrix(fx), y = sp500, nfolds = 20)

coef(lasso_model_original)
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)   0.0009599992
## australia    -0.0876032401
## brazil       -0.0667453281
## canada        .
## china         .
## denmark       .
## hong kong     .
## india         .
## japan         .
## south korea   .
## malaysia      .
## mexico       -0.5101884542
## new zealand   .
## norway        .
## singapore     .
## south africa  .
## sri lanka     .
## sweden       -0.1163433956
## switzerland   .
## taiwan        .
## thailand      .
## uk            .
## venezuela     .
## euro          .
```

```
par(mfrow = c(1,2))
plot(lasso_model, main = "LASSO PCR")
plot(lasso_model_original, main = "LASSO with Original Covariates")
```



- The plot shows the cross-validated mean squared error (CV MSE) as a function of the penalty parameter (lambda) used in the Lasso model.
- It helps in understanding how the performance of the model changes with different levels of regularization. The optimal value of lambda can be selected based on the point where CV MSE is minimized.

a. LASSO with Original Covariates:

- Direct Variable Selection: Performs direct variable selection by shrinking coefficients of less important original covariates towards zero.
- No Dimensionality Reduction: Retains the original feature space without reducing dimensionality, potentially resulting in a larger model compared to LASSO PCR.
- Interpretability: The resulting model may be more interpretable as it directly identifies the subset of original covariates that are most important for predicting the response variable.

b. LASSO PCR:

- Dimensionality Reduction: Utilizes principal components obtained through PCA, reducing the dimensionality of the feature space.
- Implicit Feature Selection: While not explicitly performing variable selection, LASSO PCR indirectly selects important features by shrinking coefficients associated with less relevant principal components.

c. Differences:

- Dimensionality vs. Direct Selection: LASSO PCR reduces dimensionality through principal components, while LASSO with Original Covariates directly selects variables.