BUSN 41201 HW7

Group 11: Yu-Ting Weng, Mengdi Hao, Elena Li, Minji Park, Sarah Lee

Question 1

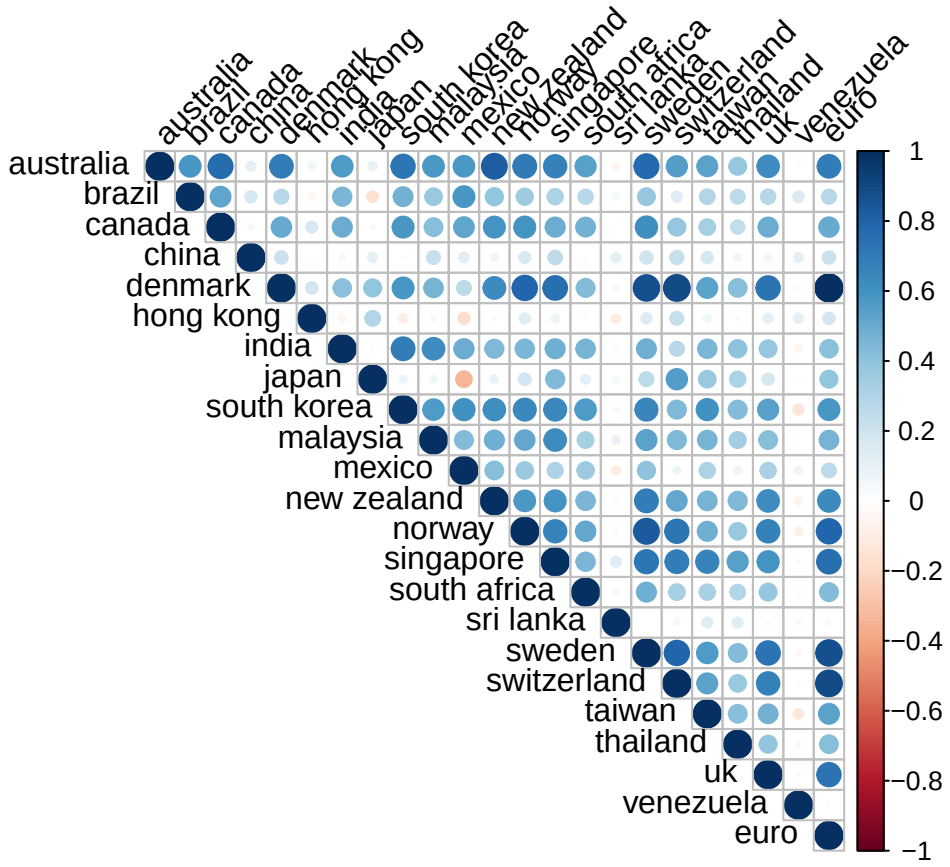
Discuss correlation amongst dimensions of fx. How does this relate to the applicability of factor modelling?

```
currency_codes <- read.table("currency_codes.txt", stringsAsFactors = FALSE)
fx_colnames <- colnames(fx)
match_currency_code <- function(colname, currency_codes) {
  code <- substr(colname, 3, 4)
  match_index <- match(code, currency_codes$V1)
  if (!is.na(match_index)) {
    return(currency_codes$V2[match_index])
  } else {
    return(colname)
  }
}

for (i in seq_along(fx_colnames)) {
  fx_colnames[i] <- match_currency_code(fx_colnames[i], currency_codes)
}
colnames(fx) <- fx_colnames
```

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

corrplot(fx_corr, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
```



We can see positive correlations between many factors in the data. High correlations among the dimensions of FX data indicate that common underlying factors may influence the exchange rates. This makes factor modeling particularly applicable, as it can identify and extract these latent factors to simplify the analysis and improve understanding of the exchange rate movements.

Question 2

Fit, plot, and interpret, principal components.

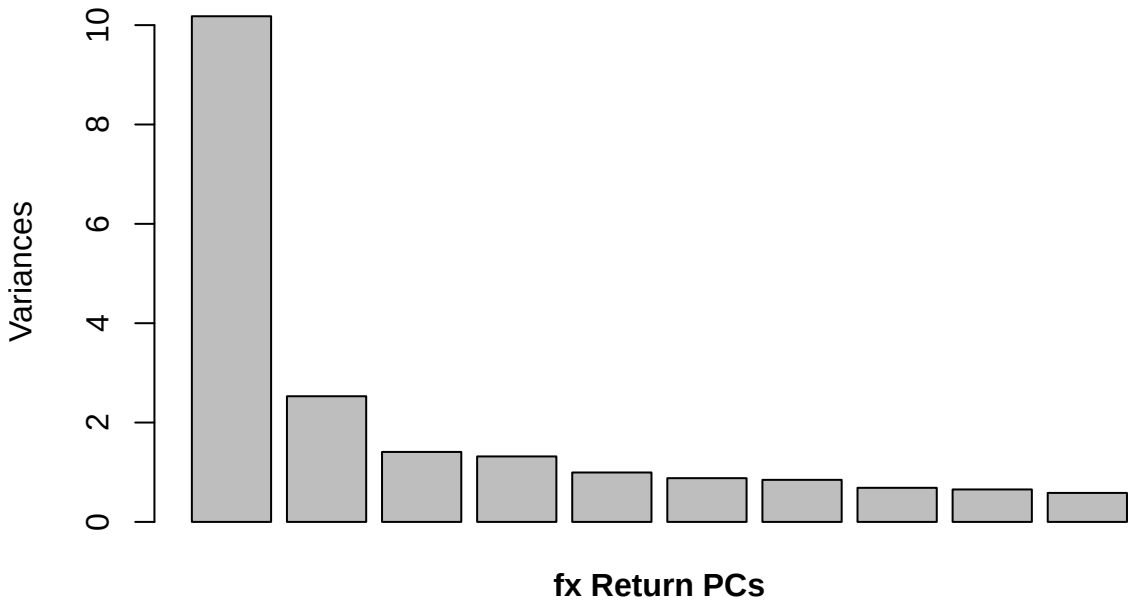
```
pca <- prcomp(fx, center = TRUE, scale = TRUE)
summary(pca)
```

```
## 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
```

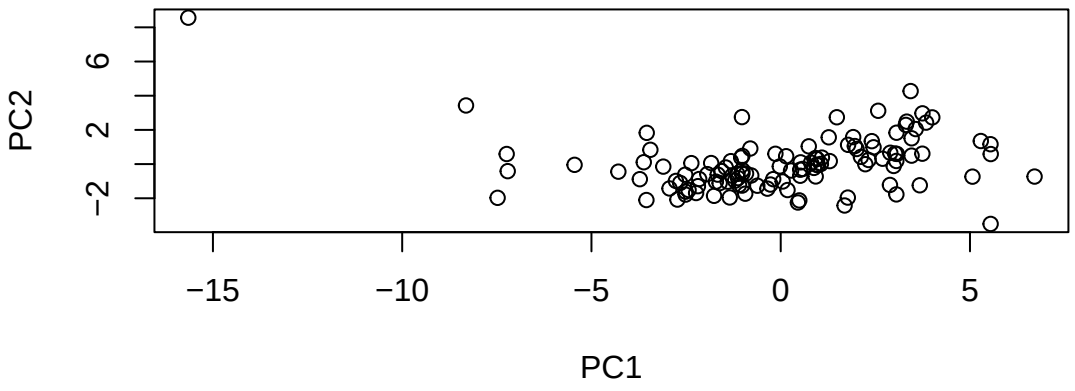
```
plot(pca, main="")

mtext(side=1, "fx Return PCs",  line=1, font=2)
```



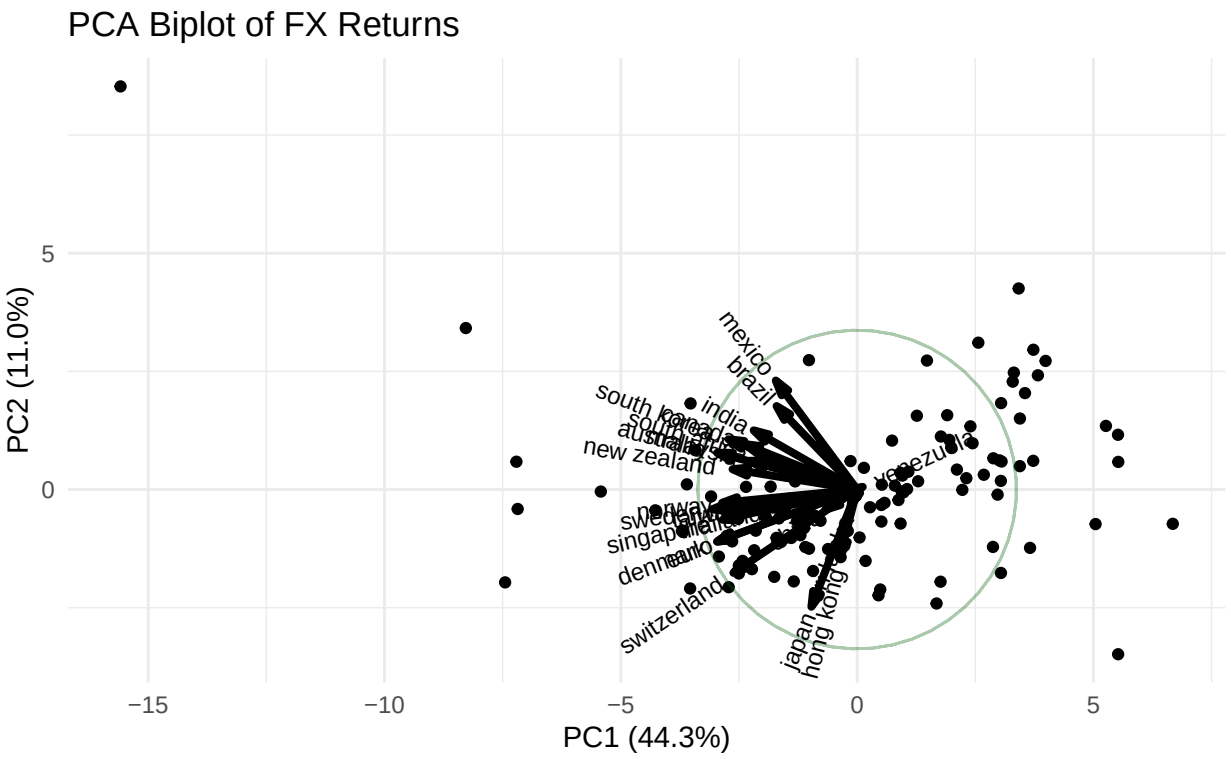
The variance plot shows that only **the first principal component significantly impacts.**

```
predict_pca <- predict(pca)
plot(predict_pca[,1:2], pch=21, bg=(4:2)[colnames(fx)], main="")
```



The scatter plot also shows that **PC1 captures most of the variance with a wide spread of data points along this axis**, while PC2 captures less variance with a narrower spread of data points. This indicates that PC1 is more influential in explaining the variability in the FX returns data.

```
ggbiplot(pca, obs.scale = 1, var.scale = 1,
         ellipse = TRUE, circle = TRUE) +
  scale_color_discrete(name = '') +
  theme_minimal() +
  ggtitle("PCA Biplot of FX Returns") +
  theme(legend.position = "bottom")
```



The biplot shows that the first two principal components explain a significant portion of the variance in the FX returns data, with PC1 accounting for 44.3% and PC2 accounting for 11.0% of the variance. It also reveals that Mexico, Switzerland, and Japan have the largest influence on these components, indicating their strong contribution to the variability in the data.

```
t(round(pca$rotation[,1:2],2))
```

```
##     australia brazil canada china denmark hong kong india japan south korea
## PC1     -0.28  -0.16  -0.22 -0.06   -0.28     -0.03 -0.20 -0.09       -0.25
## PC2      0.14   0.33   0.18 -0.08   -0.21     -0.26  0.23 -0.46        0.20
##     malaysia mexico new zealand norway singapore south africa sri lanka sweden
## PC1    -0.21  -0.16       -0.25  -0.27     -0.26        -0.19     -0.01  -0.29
## PC2     0.11   0.43        0.08  -0.05     -0.12         0.12     -0.06  -0.08
##     switzerland taiwan thailand    uk venezuela  euro
## PC1       -0.24  -0.21    -0.17 -0.24      0.01 -0.28
## PC2       -0.33  -0.07    -0.08 -0.08      0.01 -0.21
```

The loadings for the first principal component (PC1) show that Australia, Denmark, Hong Kong, and Switzerland have the highest negative contributions, indicating that these currencies are the most influential in explaining the variance captured by PC1. For the second principal component (PC2), Japan, Brazil, and Mexico have the highest absolute loadings, suggesting they play a significant role in the variance captured by PC2.

```
sort(predict_pca[,1])[1:5]
```

```
##    OCT2008    SEP2008    AUG2008    NOV2008    MAY2010
## -15.650761  -8.314062  -7.479007  -7.239428  -7.212144
```

```
sort(predict_pca[,1], decreasing = TRUE)[1:5]
```

```
##  MAY2009  MAY2003  APR2009  OCT2010  NOV2004
## 6.706182 5.548390 5.545457 5.542401 5.286064
```

```
sort(predict_pca[,2])[1:5]
```

```
##   APR2009   APR2010   APR2003   MAY2007   JUN2005
## -3.499584 -2.421432 -2.247858 -2.123700 -2.102884
```

1. The first result indicates that October 2008, September 2008, August 2008, November 2008, and May 2010 have the lowest scores on PC1, suggesting that these months contributed the most negatively to the first principal component.
2. The second result shows that May 2009, May 2003, April 2009, October 2010, and November 2004 have the highest scores on PC1, indicating that these months contributed the most positively to the first principal component.
3. The third result reveals that April 2009, April 2010, April 2003, May 2007, and June 2005 have the lowest scores on PC2, indicating that these months contributed the most negatively to the second principal component.

Question 3

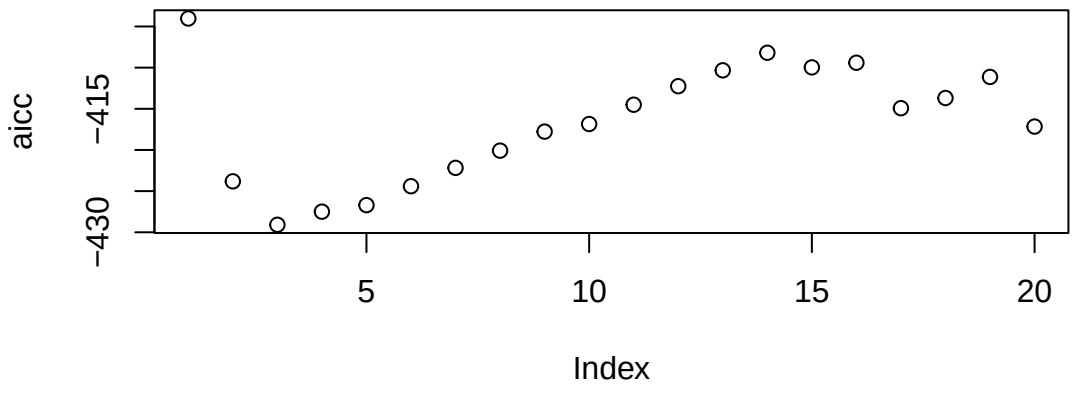
Regress SP500 returns onto currency movement factors, using both 'glm on first K' and lasso techniques. Use the results to add to your factor interpretation.

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

```
zdf <- as.data.frame(predict_pca)

kfits <- lapply(1:20,
    function(K) glm(sp500~., data=zdf[,1:K,drop=FALSE]))

aicc <- sapply(kfits, AICc) # apply AICc to each fit
plot(aicc)
```



```
which.min(aicc) ## it likes 3 factors best
```

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

We can see that AICc building one-at-a-time chooses K=3.

```
summary(spglm <- glm(sp500 ~ ., data=zdf[,1:3]))
```

```
##
## Call:
## glm(formula = sp500 ~ ., data = zdf[, 1:3])
##
## 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
```

- Lasso

```
lassoPCR <- cv.gamlr(x=predict_pca, y=sp500, nfold=20)
coef(lassoPCR)
```

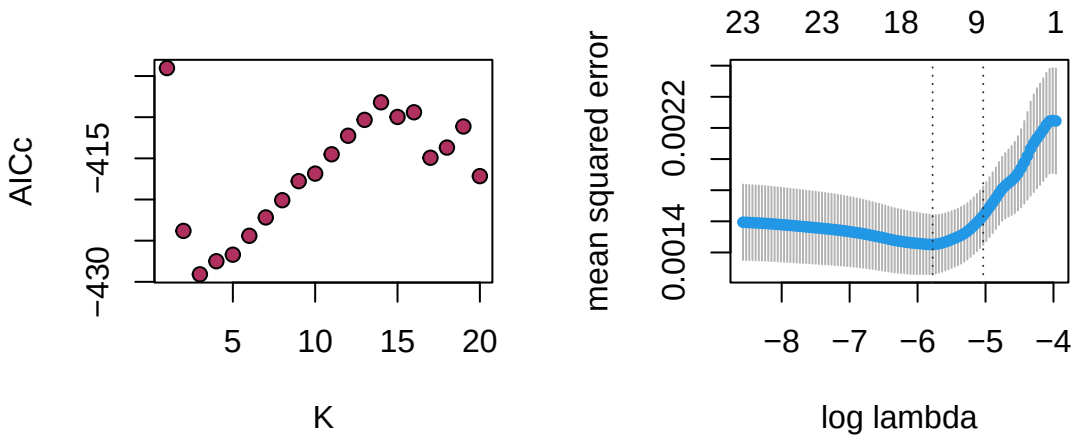
```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                    seg24
## intercept  0.0004430924
## PC1        0.0039247159
## PC2       -0.0070684101
## PC3       -0.0027007248
## PC4        .
## PC5        .
## PC6        .
## PC7        .
## PC8        .
## PC9        .
## PC10       .
## PC11       .
## PC12       .
## PC13       .
## PC14       .
## PC15      -0.0012033897
## PC16       .
## PC17      -0.0065582481
## PC18       .
## PC19       .
```

```
## PC20      -0.0085143359
## PC21       .
## PC22       .
## PC23       0.1431123816
```

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

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

plot(lassoPCR)
```



From the AICc results, we determined that K=3, indicating the first three principal components (PC1, PC2, PC3), are the most significant in explaining the variance in S&P 500 returns. However, the CV Lasso results highlighted additional factors (PC15, PC17, PC20, PC21, PC23), suggesting that these components also significantly explain the variance. Therefore, our factor interpretation reveals that while PC1, PC2, and PC3 are primary drivers, other components like PC15, PC17, PC20, PC21, and PC23 capture additional variations essential for a more comprehensive model.

Question 4

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

- Fit Lasso to the original covariates

```
fx_returns_scaled <- scale(fx)
cv_lasso_original <- cv.glmnet(as.matrix(fx_returns_scaled), sp500, alpha = 1, nfolds = 20)

best_lambda_original <- cv_lasso_original$lambda.min

lasso_original_model <- glmnet(as.matrix(fx_returns_scaled), sp500, alpha = 1, lambda = best_lambda_ori

lasso_original_coef <- coef(lasso_original_model)
print("Coefficients of Lasso model with original covariates:")
```

```
## [1] "Coefficients of Lasso model with original covariates:"
```

```
print(lasso_original_coef)
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)   0.0004430924
## australia    -0.0080566572
## brazil       -0.0060148618
## canada        0.0029900393
## china         0.0059594305
## denmark       .
## hong kong    -0.0032118514
## india        -0.0006254756
## japan         0.0029733982
## south korea   .
## malaysia     -0.0019937631
## mexico       -0.0146034311
## new zealand   0.0029386448
## norway        0.0114248568
## singapore     0.0072260547
## south africa -0.0051222560
## sri lanka     0.0005766097
## sweden       -0.0263966251
## switzerland   .
## taiwan        .
## thailand      .
## uk            0.0070977949
## venezuela    -0.0044995369
## euro          .
```

- PCR

```
cv_lasso_pca <- cv.glmnet(as.matrix(zdf), sp500, alpha = 1, nfolds = 20)

best_lambda_pca <- cv_lasso_pca$lambda.min

lasso_pca_model <- glmnet(as.matrix(zdf), sp500, alpha = 1, lambda = best_lambda_pca)

lasso_pca_coef <- coef(lasso_pca_model)
print("Coefficients of Lasso PCR model:")
```

```
## [1] "Coefficients of Lasso PCR model:"
```

```
print(lasso_pca_coef)
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)  0.0004430924
## PC1          0.0051273031
## PC2         -0.0094807689
## PC3         -0.0059335652
## PC4          0.0000727116
## PC5         -0.0015303719
## PC6          .
## PC7          .
## PC8          .
## PC9          .
## PC10        -0.0021174842
## PC11         .
```

```
## PC12         .
## PC13        -0.0002729995
## PC14         .
## PC15        -0.0081471935
## PC16         0.0045376558
## PC17        -0.0151561772
## PC18        -0.0033998940
## PC19         .
## PC20        -0.0198913286
## PC21        -0.0120392083
## PC22         .
## PC23         0.3895367690
```

- Comparison

```
# Compare the number of selected variables
num_nonzero_original <- sum(lasso_original_coef != 0) - 1  # subtract 1 for intercept
num_nonzero_pca <- sum(lasso_pca_coef != 0) - 1  # subtract 1 for intercept

cat("Number of selected variables in original covariates:", num_nonzero_original, "\n")
```

```
## Number of selected variables in original covariates: 17
```

```
cat("Number of selected variables in PCR:", num_nonzero_pca, "\n")
```

```
## Number of selected variables in PCR: 14
```

We can see that PCR reduced the number of selected variables from 15 in the original covariates to 12, demonstrating its effectiveness in dimension reduction. Additionally, because the principal components are orthogonal and independent, they eliminate multicollinearity issues, leading to more stable and reliable regression estimates.