GEOG 714 - Assignment 7

Haoran Xu

2024-11-10

```
# libraries
library(here)
```

```
## here() starts at /Users/horanxu/Desktop/GEOG714_Applied_Data_Analysis_for_Geographers_and_Earth_Scien
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
df1 <- data.frame(read.csv(paste0(here(), "/Assignments/Assignment7/Greece_precip_data.csv")))
```

```
corr <- as.data.frame(cor(df1[,4:15]))
```

Q1. Do some simple descriptive analysis, and provide a few brief sentences that summarise the information in this table.

```
summary(df1)
summary(corr)

mean_corr <- mean(corr[lower.tri(corr)])

corrplot(as.matrix(corr), method = "circle", type = "lower",
         title = "Monthly Precipitation Correlations", tl.cex = 0.8)
```

The average precipitation first decreases and then increased from Jan to Dec in Greece, with the lowest in August and highest in Dec. The correlation matrix showed similar pattern that the average correlation between two months for a month generally decreases and then increases from Jan to Dec, ranging from 0.489 to 0.831, with the mean correlations of all being 0.738, suggesting rather significant positive correlation. From the correlation plot, a seasonal structure can be shown, where months within the same season (e.g., winter or spring) have higher correlations with each other, while months from different seasons (e.g., winter vs. summer) tend to have lower correlations.

```
pca_out <- prcomp(df1[, 4:15],
                  center=TRUE, scale=TRUE)
# scale=TRUE would mean the standarlizaiton
summary(pca_out)

plot(pca_out$x[, 1], pca_out$x[, 2]) # scores for each 75 points on either PC1 or PC2
plot(df1$LON, df1$LAT, col="white")
text(df1$LON, df1$LAT, labels=df1$ID)
```

Q2. Remove the observations from stations #1 and #37 and re-run the PCA analysis. Comment briefly on how this has changed the results.

```
df2 <- df1[df1$ID != 1 & df1$ID != 37,]
pca_out2 <- prcomp(df2[, 4:15],
                   center=TRUE, scale=TRUE)
# scale=TRUE would mean the standarlizaiton
summary(pca_out2)

plot(pca_out2$x[, 1], pca_out2$x[, 2]) # scores for each 75 points on either PC1 or PC2
```

The portion of variances explained of PC1 has increased slightly from 77.17% to 79.06%. Yet the portions of PC2 and PC3 have decreased slightly. And the new scores plot of PC1 and PC2 does not have two previous extreme values which are situated in the “high and low” end of PC2. This conforms to the previous PC portion of variances explained.

```
df2$station_sum <- apply(df2[, 4:15], 1, function(x) sum(x))
month_totals <- apply(df2[, 4:15], 2, function(x) sum(x))
```

```
set.seed(8675309)
x1 <- c(0)
x2 <- c(0)
x3 <- c(0)
x4 <- c(0)
x5 <- c(0)
x6 <- c(0)
for(i in 1:300){
  x1[i] <- rnorm(1)
  x2[i] <- rnorm(1) + x1[i]
  x3[i] <- 2*rnorm(1) + x1[i] + x2[i]/2
  x4[i] <- rnorm(1) + x2[i] + x3[i]/2
  x5[i] <- rnorm(1) - x1[i]
  x6[i] <- rnorm(1) + x1[i]/3
}

m <- data.frame(cbind(x1, x2, x3, x4, x5, x6))
```

Q3. Create a correlation matrix using all the variables created in the previous step. Based on the code above and the results from the correlation matrix, discuss briefly what you expect to see from PCA. How many principal components might you expect to see explaining the bulk of variability in these data? Explain your answer in two to three sentences.

```
corr2 <- cor(m)
```

I expect to see in the PCA that the first component would not be enough to explain the variability very well, as `x3` and `x4` are not directly correlated to `x1` and `x5` has negative correlation and that `x6` has minor correlations with `x1`. Variable `x3` has to do with x1 and x2 (the latter of which is mainly correlated with `x1`), and variable `x4` has to do with `x2` and `x3`, making it having more indirect correlation with `x1`. I would expect **2 or 3** principal components explaining the bulk of variability.

```r
pca_out3 <- prcomp(m,
                   center=TRUE, scale=TRUE)

summary(pca_out3)
```

Q4. Create a biplot of these data, and then using the biplot and the results of the table of principal components, provide an interpretation of the structure of these data. Consider the angles of the red lines of the biplot. Are they pointing in the same direction? Are they at right angles? Are they pointing in opposite directions? Look back at how the data are synthesised and discuss what might explain the angle that these arrows are pointing. Explain your answer in no more than 3 to 4 sentences.

```r
biplot(prcomp(m, center=TRUE, scale=TRUE), main = "PC1 and PC2")
```

```r
biplot(prcomp(m, center=TRUE, scale=TRUE), choices = 2:3, main = "PC2 and PC3")
```

The eigenvectors of `x1`, `x2`, `x3`, `x4`, `x5` can be mostly explained by PC1 as they parallel well with the PC1 axis, with all but `x5` explained positively by PC1. The `x6` is more explained by PC2 as it is inclined to parallel with the PC2 axis. The reasons for the biplot patterns are that `x5` has negative correlation with `x1`; `x3` and `x4` are more indirectly correlated with `x1`; and that `x6` has minor correlations with `x1` becasue it has large random variances.

```r
var = pca_out3$sdev ^ 2 / sum(pca_out3$sdev ^ 2)
plot(var, xlab="Principal Component", ylab="Variance Explained")
```

Q5. Modify the code for this plot in some way to make it look nice and stuff!

```r
plot(var, type = "b", pch = 19, col = "blue", lwd = 2,
     xlab = "Principal Component", ylab = "Variance Explained",
     main = "Variance Explained by Each Principal Component",
     ylim = c(0, 1))

cumulative_var <- cumsum(var)
lines(cumulative_var, type = "b", pch = 17, col = "red", lwd = 2)

legend("right", legend = c("Variance Explained", "Cumulative Variance"),
       col = c("blue", "red"), pch = c(19, 17), lwd = 2)

text(1:length(var), var, labels = round(var, 3), pos = 3, cex = 0.8, col = "blue")
```

```r
set.seed(911)
v1 <- rnorm(1000)
v2 <- rnorm(1000)
v3 <- rnorm(1000) + v1
v4 <- 4 * rnorm(1000) - v1
```

Q6. Write code that measures the correlations between all these variables.

```r
m2 <- data.frame(cbind(v1, v2, v3, v4))
corr3 <- cor(m2)
```

```r
pca_out4 <- prcomp(cbind(v1, v2, v3, v4),
                   center=TRUE, scale=TRUE)
summary(pca_out4)

plot(pca_out4$x[, 1], pca_out4$x[, 2],
     main="Score plot", pch=20,
     xlab="PC1", ylab="PC2"
)
```

```r
biplot(prcomp(m2, center=TRUE, scale=TRUE), main = "PC1 and PC2")
```

```r
biplot(prcomp(m2, center=TRUE, scale=TRUE), choices = c(1, 3), main = "PC1 and PC3")
```

```r
biplot(prcomp(m2, center=TRUE, scale=TRUE), choices = 2:3, main = "PC2 and PC3")
```

Q7. Reflect on why you see this pattern on the plot, and alter the process for creating synthetic data to generate a pattern with some sort of clustering or clumping of two or more groups in the PC1 and PC2 space. This may require some experimentation/trial and error. Hint: think about changing the means and/or variances of the random number generating function for two of the variables. If subsets of the data have different means and/or variances, this can produce 'clustering'. Don't rush this step of the experimentation process, as this exploration process will give you a deeper understanding of what the underlying data are saying. Once you are done, re-run the prcomp() step and plot the components to show the 'clumping' with the synthetic data you create. Plot the biplot as well.

```r
set.seed(911)

v1 <- rnorm(500)
v2 <- rnorm(500)
v3 <- rnorm(500) + v1
v4 <- 4 * rnorm(500) - v1

v1_2 <- rnorm(500, mean = 3)
```

```
v2_2 <- rnorm(500, mean = -2)
v3_2 <- rnorm(500, mean = 2) + v1_2
v4_2 <- 4 * rnorm(500) - v1_2

v1 <- c(v1, v1_2)
v2 <- c(v2, v2_2)
v3 <- c(v3, v3_2)
v4 <- c(v4, v4_2)

m3 <- data.frame(cbind(v1, v2, v3, v4))

corr3 <- cor(m3)

pca_out5 <- prcomp(m3, center = TRUE, scale = TRUE)
summary(pca_out5)

plot(pca_out5$x[, 1], pca_out5$x[, 2],
     main = "Score Plot with Clustering",
     pch = 20, col = rep(c("blue", "red"), each = 500),
     xlab = "PC1", ylab = "PC2")
```

```
biplot(pca_out5, main = "PC1 and PC2 with Clustering")
```