Lab 13. PCA and MCA

```
library(tidyverse)
library(ggfortify)
```

1 Gospels' frequency word lists

The gospels of Matthew, Mark, and Luke are referred to as the Synoptic Gospels and stand in contrast to John, whose content is comparatively distinct. This dataset (https://tinyurl.com/y8tcf3uw) contains frequency of selected words (without stopwords, without pronouns and without frequent word "Jesus") as attested in four gospels of the New Testament.

For some visualisations you will need assign row names to the dataframe:

```
gospels <- read.csv("https://tinyurl.com/y8tcf3uw")
row.names(gospels) <- gospels$word
```

1.1 Apply PCA to four continuous variables. What is the cumulative proportion of explained variance for the first and second component?

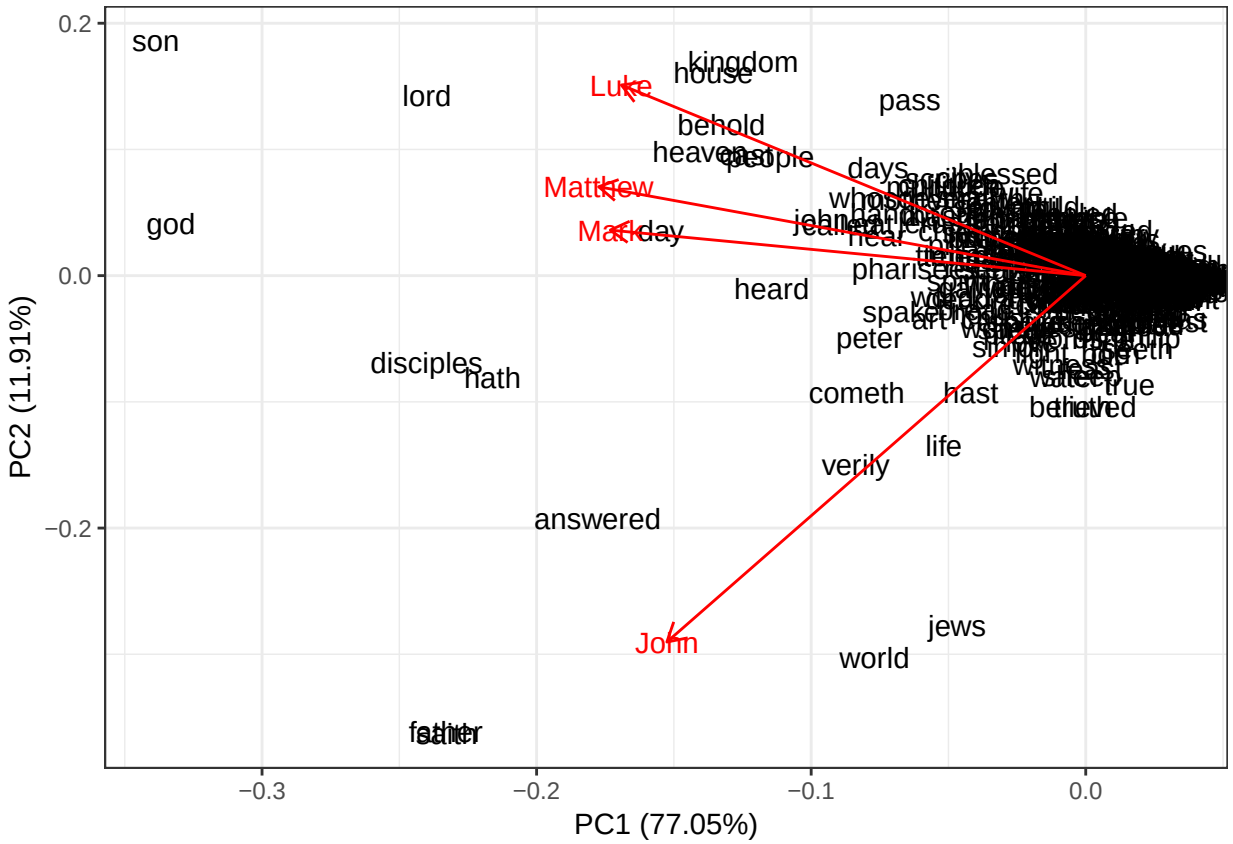
```
PCA <- prcomp(gospels[,2:5], center = TRUE, scale. = TRUE)
summary(PCA)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7556 0.6903 0.50983 0.42619
## Proportion of Variance 0.7705 0.1191 0.06498 0.04541
## Cumulative Proportion  0.7705 0.8896 0.95459 1.00000
```

1.2 Use the `autoplot()` function of the library ggfortify for creating plot like this.

See more examples here: https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

```
autoplot(PCA,
         shape = FALSE,
         loadings = TRUE,
         label = TRUE,
         loadings.label = TRUE)+
  theme_bw()
```



1.3 Predict the coordinates for the word "Jesus", which have the following frequencies: John = 0.05, Luke = 0.01, Mark = 0.02, Matthew = 0.02.

```
predict(PCA, data.frame(John = 0.05, Luke = 0.01, Mark = 0.02, Matthew = 0.02))
```

```
##            PC1       PC2      PC3      PC4
## [1,] -22.60497 -9.599171 2.367918 2.104944
```

2. Register variation in the British National Corpus

Dataset and discription from Natalia Levshina's package Rling.

This is a data set with relative frequencies (proportions) of different word classes in 69 subcorpora of the British National Corpus (the BYU-BNC version). Reg — a factor that describes the metaregister with levels Acad, Fiction, Misc, News, NonacProse and Spok Ncomm — a numeric vector with relative frequencies of common nouns. Nprop — a numeric vector with relative frequencies of proper nouns. Vpres — a numeric vector with relative frequencies of verbs in the present tense form, 3rd person singular. Vpast — a numeric vector with relative frequencies of verbs in the past tense form. P1 — a numeric vector with relative frequencies of the first-person pronouns. P2 — a numeric vector with relative frequencies of the second-person pronouns. Adj — a numeric vector with relative frequencies of adjectives. ConjCoord — a numeric vector with relative frequencies of coordinating conjunctions. ConjSub — a numeric vectorwith relative frequencies of subordinating conjunctions. Interject — a numeric vector with relative frequencies of interjections. Num — a numeric vector with relative frequencies of numerals.

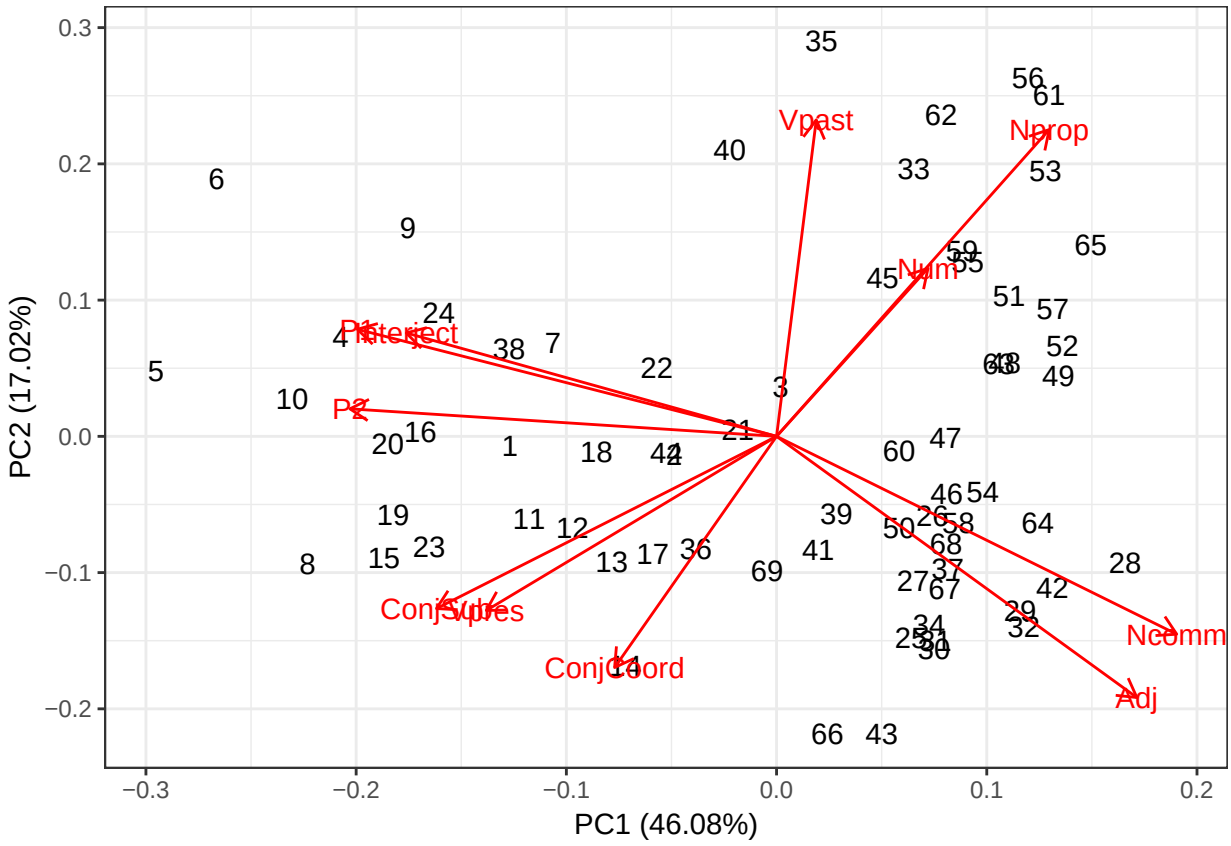
2.1 Apply PCA to all variables. What is the cumulative proportion of explained variance for the first, second and third components?

2.2 Extract the coordinates from the pca object (pca$x), merge with the dataset itself, and create a visualization using the first two components and creating confidence ellipses for each metaregister.

```
reg_bnc <- read.csv("https://goo.gl/19QywL")
pca <- prcomp(reg_bnc[,-1], center = TRUE, scale. = TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6    PC7
## Standard deviation     2.2513 1.3683 1.1730 0.88886 0.80320 0.64940 0.5480
## Proportion of Variance 0.4607 0.1702 0.1251 0.07183 0.05865 0.03834 0.0273
## Cumulative Proportion  0.4607 0.6310 0.7560 0.82786 0.88650 0.92484 0.9521
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.43204 0.37908 0.32981 0.29551
## Proportion of Variance 0.01697 0.01306 0.00989 0.00794
## Cumulative Proportion  0.96911 0.98217 0.99206 1.00000
```

```
autoplot(pca,
         shape = FALSE,
         loadings = TRUE,
         label = TRUE,
         loadings.label = TRUE)+
  theme_bw()
```




```
reg_bnc <- cbind(reg_bnc, pca$x)

reg_bnc %>%
  ggplot(aes(PC1, PC2, color = Reg))+
```

```
  geom_point()+
  stat_ellipse()+
  theme_bw()
```

```
## Warning: Removed 1 rows containing missing values (geom_path).
```