Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering
IDP Institut für Datenanalyse
und Prozessdesign

HS 2016

Statistisches Data Mining (StDM)

Woche 1

Aufgabe 1 PCA in R

Arbeiten Sie das Kapitel 10.4 Lab 1 Principal Components Analysis in ISLR durch.

Aufgabe 2 Abstimmungsverhalten der Kantone

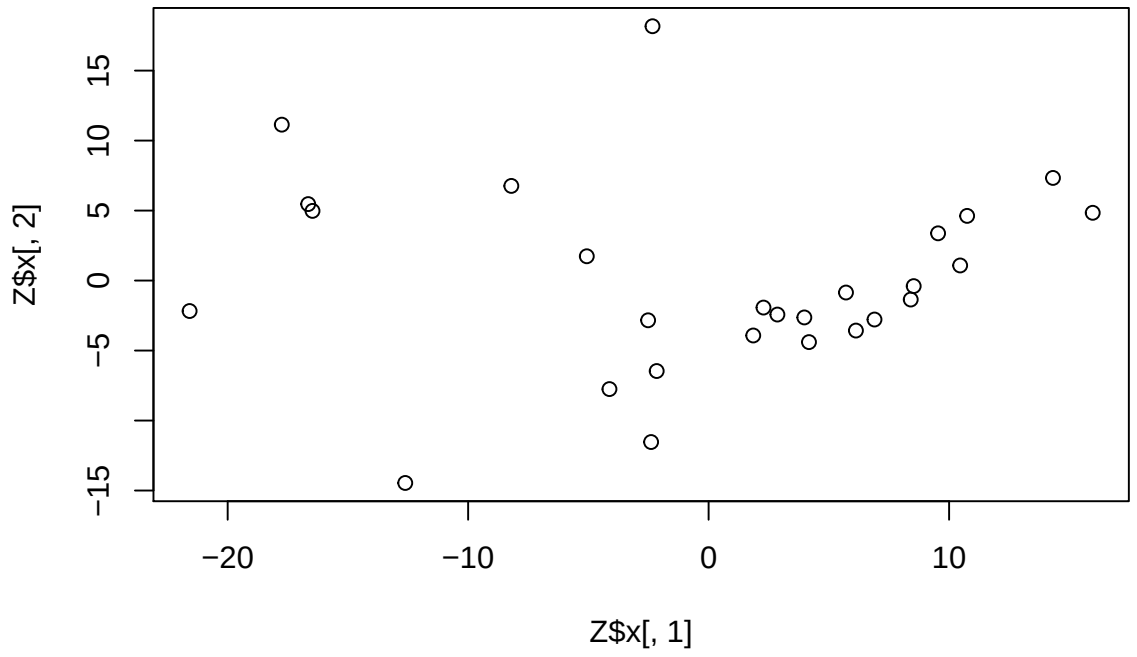
Besorgen Sie sich den Datensatz der Volksabstimmungen in der Schweiz von der Webseite: http://www.bfs.admin.ch/bfs/portal/de/index/themen/17/03/blank/data/01.html dort unter `Online-Datenrecherche Volksabstimmungen (Ergebnisse Ebene Kanton seit 1866)`. Wählen Sie alle Kantone und alle Vorlagen ab 1990 aus. Wählen Sie als Ergebniss das Feld `Ja in \%`. Alternativ können Sie auch den Datensatz 'px-x-1703030000_100.csv' verwenden, der mit den oben beschriebenen Methoden erzeugt worden ist.

a) Laden Sie den Datensatz und bringen Ihn mit der Funktion `reshape` in eine Matrixform, bei der die Kantone die Zeilen und die Abstimmungsergebnisse die Spalten sind.

```
X.t <- read.table(file.path(baseDir, "px-x-1703030000_100.csv"), header=T, sep=";", skip=1, string
X <- reshape(X.t, idvar = "Kanton", timevar = c("Datum.und.Vorlage"), direction = "wide")
names = X$Kanton
X = data.matrix(X[,2:ncol(X)])
```

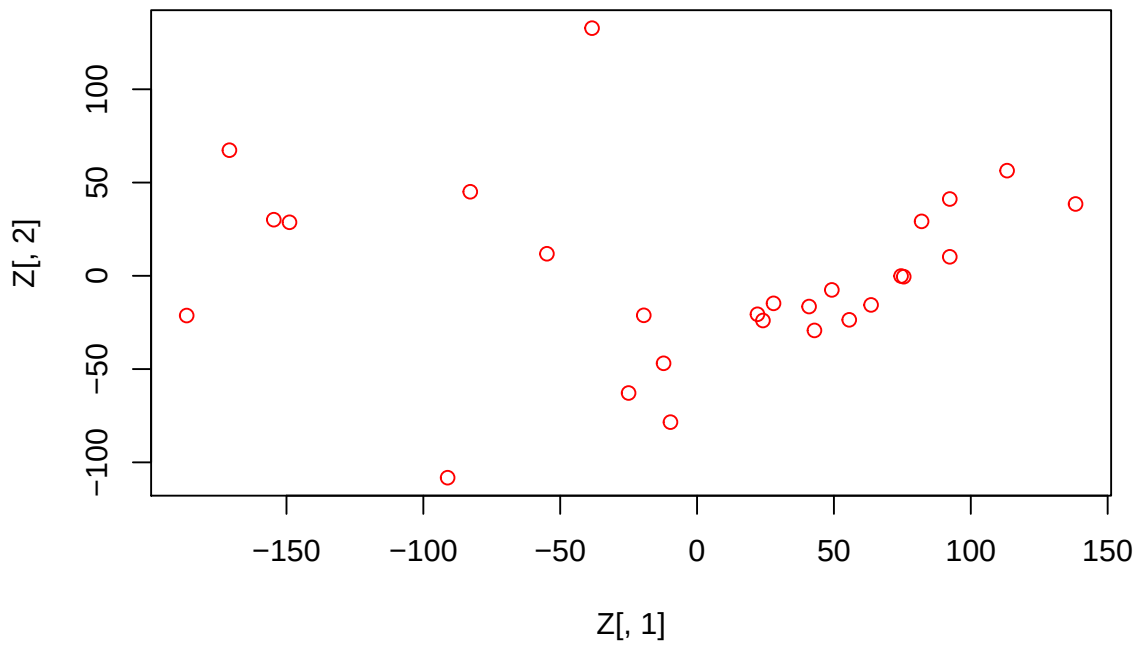
b) Verschaffen Sie sich einen Überblick über die Daten in dem Sie eine PCA durchführen, plotten Sie auch die Kantonsnamen.

```
Z <- prcomp(X, scale. = TRUE)
plot(Z$x[,1], Z$x[,2])
```

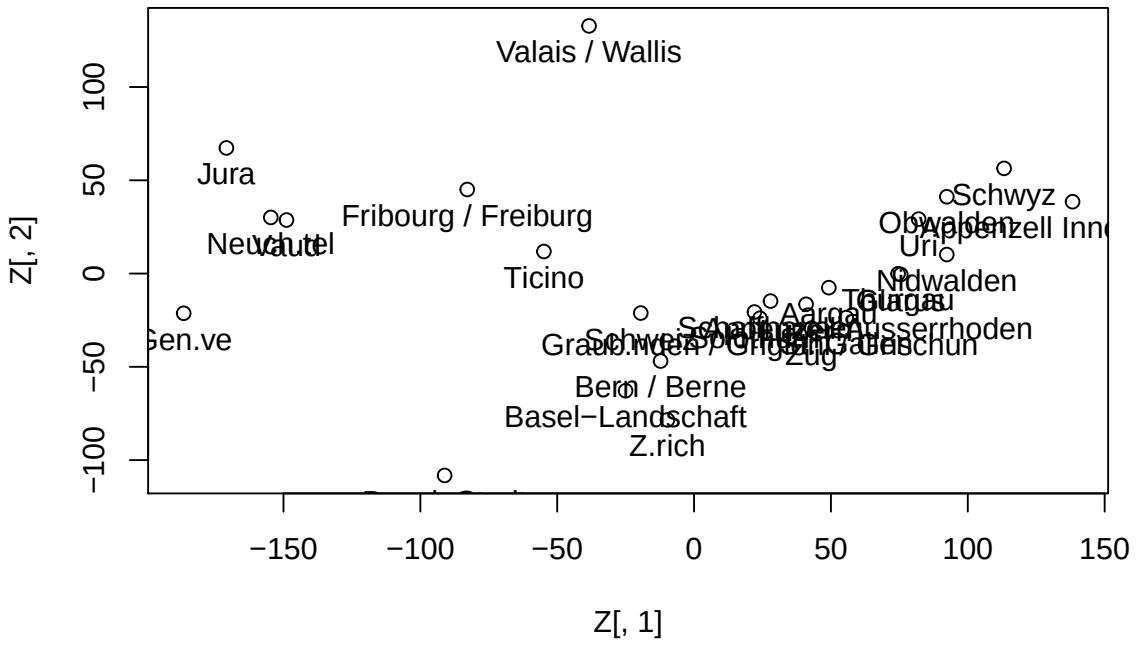



```
#text(abst.pcS$x, labels = names)

Z <- prcomp(X, scale. = FALSE)$x #OK ohne Skalierung, da Prozente. Bis auf Skala gibt es keinen Un
plot(Z[,1], Z[,2], col='red')
```

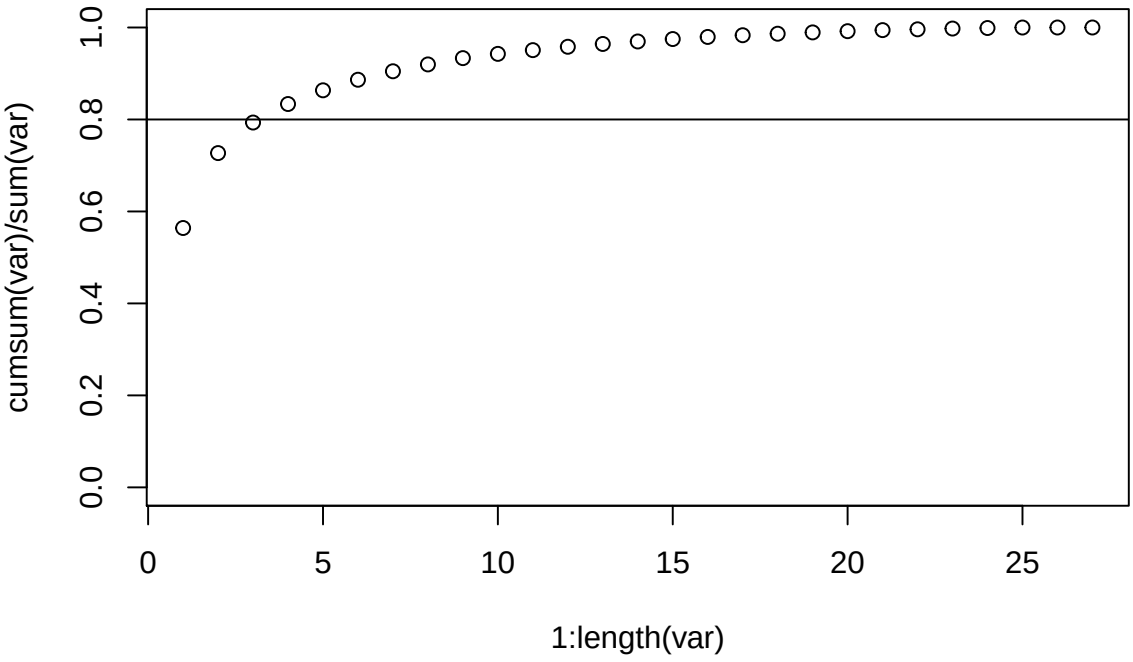



```
plot(Z[,1], Z[,2])
text(Z, labels = names,pos=1)
```

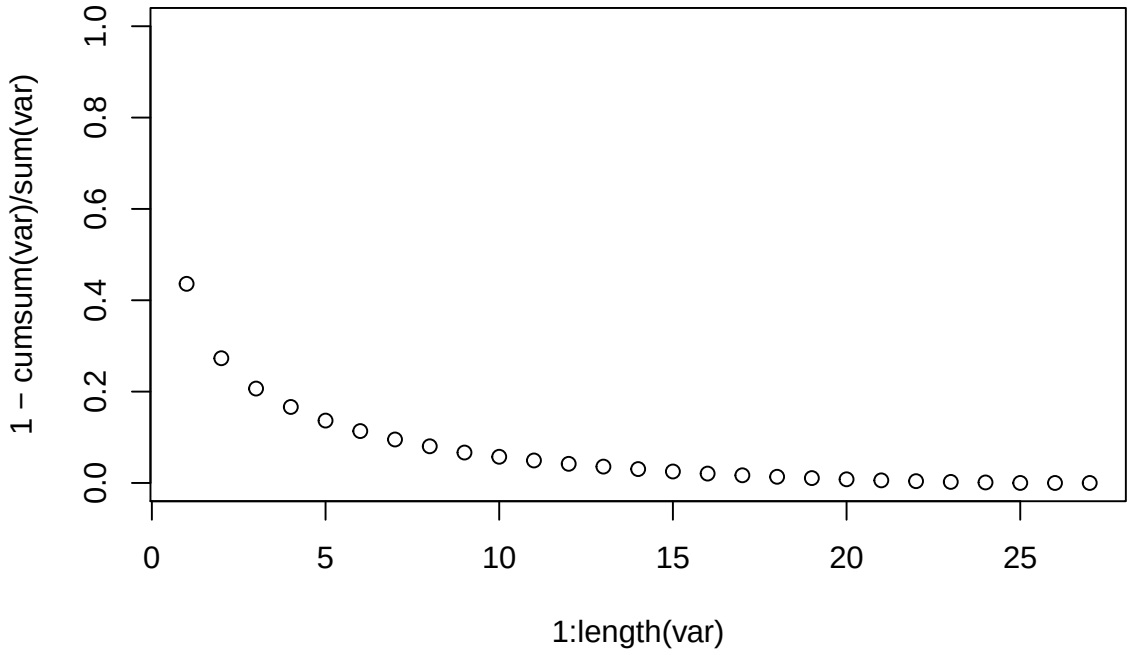


c) Ist eine zweidimensionale Darstellung angebracht? Ploten Sie dazu die `explained variance` gegen die Anzahl der Hauptkomponenten.

```
abst.pcS <- prcomp(X, scale. = FALSE)
var = abst.pcS$sdev^2
plot(1:length(var), cumsum(var)/sum(var), ylim=c(0,1))
abline(h=0.8)
```




```
plot(1:length(var), 1 -  cumsum(var)/sum(var), ylim=c(0,1))
```



Aufgabe 3 Wordembedding

Visualization of a word-embedding: The data-set (`samplewordembedding.csv`) has been produced by the word2vec algorithm (see here for details.) Each of the 871 word in the corpus is assigned to a 300 dimensional vector.

a) Load the data set and perform a PCA.

```
wrds = read.csv(file.path(baseDir, 'samplewordembedding.csv'))
dim(wrds)
set.seed(1)
pca = prcomp(wrds)
plot(pca$x[,1], pca$x[,2], pch='.')
idx = sample(1:nrow(wrds), 50)
text(pca$x[idx,1], pca$x[idx,2], labels=row.names(wrds)[idx],pos=1)
```

b) Evaluate the quality of the 2 dimnesional PCA plot by showing the explained variance.

```
abst.pcS <- prcomp(wrds, scale. = FALSE)
var = abst.pcS$sdev^2
plot(1:length(var), cumsum(var)/sum(var), ylim=c(0,1), type='l')
plot(1:length(var), 1 -  cumsum(var)/sum(var), type='l'))
abline(h=0.8)
```