Solution Lesson01: PCA

Machine Learning using R

Exercise 1: Wisconsin breast cancer data

The Wisconsin data set describes features computed from digitized images of fine needle asprirates (FNA) of breast mass. In addition to the extracted features, the data set includes an ID number of the image and the diagnosis of the sample (M = malignant, B = benign). For more information see https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

a) Load the data "breastCancer_Wisconsin.csv" using the `read.csv('breastCancer_Wisconsin.csv', stringsAsFactors=TRUE)` command. The data file can be found in the folder of this exercise. Is there a variable which is inappriopriately coded? Recode the variable into the correct format.

```
# Load the data
dat <- read.csv('breastCancer_Wisconsin.csv', header = TRUE, stringsAsFactors = TRUE)
str(dat)
```

```
## 'data.frame':    569 obs. of  32 variables:
##  $ id                      : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844
##  $ diagnosis               : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ radius_mean             : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean            : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean          : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean               : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean         : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean        : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean          : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points_mean     : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean           : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean  : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se               : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se              : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se            : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                 : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se           : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se          : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se            : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points_se       : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se             : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se    : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst            : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst           : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst         : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst              : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst        : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst       : num  0.666 0.187 0.424 0.866 0.205 ...
```

```
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

```
dat$id <- as.factor(dat$id)
```

The picuture-ID (`id`) is coded as an integer. While this will not affect the work in this series, it is usually inadvisable to code an ID-variable as a numeric. We can turn the variable into a factor using `dat$id <- as.factor(dat$id)`.

b) How many variables are there in total in the data? How many observations?

```
dim(dat)
```

```
## [1] 569  32
```

There are 32 variables (columns) in total (including the `id` and the `diagnosis` factors). There are 569 observations (rows).

c) Perform a Principal Component Analysis (with scaling) on the data. Do not include the `id` factor in the PCA. Also exclude the `diagnosis` factor from the PCA. Why can't we include these factors in the PCA? (**Hint**: `prcomp(..., scale.=TRUE)`)

```
pca_obj <- prcomp(dat[,-c(1,2)], scale. = TRUE)
```

PCA only works with numerical values, therefore factors cannot be included. Also, the ID of the pictures is in itself not a meaningful variable (regarding the picture features) and therefore it doesn't make sense to include it.

d) Check the results of the PCA. What proportion of the variance is explained by the first PC? What proportion of the variance is explained by the first four PCs together?

```
summary(pca_obj)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172
## Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251
## Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010
##                           PC8    PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.39624
## Proportion of Variance 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.00523
## Cumulative Proportion  0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.98335
##                           PC15    PC16    PC17    PC18    PC19    PC20   PC21
## Standard deviation     0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1731
## Proportion of Variance 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010
## Cumulative Proportion  0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9966
##                           PC22    PC23   PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987
```

```
## Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005
## Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
##                            PC29    PC30
## Standard deviation     0.02736 0.01153
## Proportion of Variance 0.00002 0.00000
## Cumulative Proportion  1.00000 1.00000
```

The first PC explaines a proportion of 0.4427 of the variance. The first four PCs cummulatively explain a proportion of 0.7924 of the total variance.

e) Which of the variables contributes mostly to the first PC (in absolute terms)? How large is this loading? (**Hint**: `which.max()`, `abs()`, `pca_object$rotation`)

```
(ind.m <- which.max(abs(pca_obj$rotation[,"PC1"]))) # find the index of the maximum loading
```

```
## concave.points_mean
##                   8
```

```
# Putting a value assignment command (<-) in brackets makes R show the assigned value

pca_obj$rotation[ind.m,"PC1"] # show the loading
```

```
## [1] -0.2608538
```

The variable with the highest loading is `concave.points_mean`, it has a loading of -0.2609.

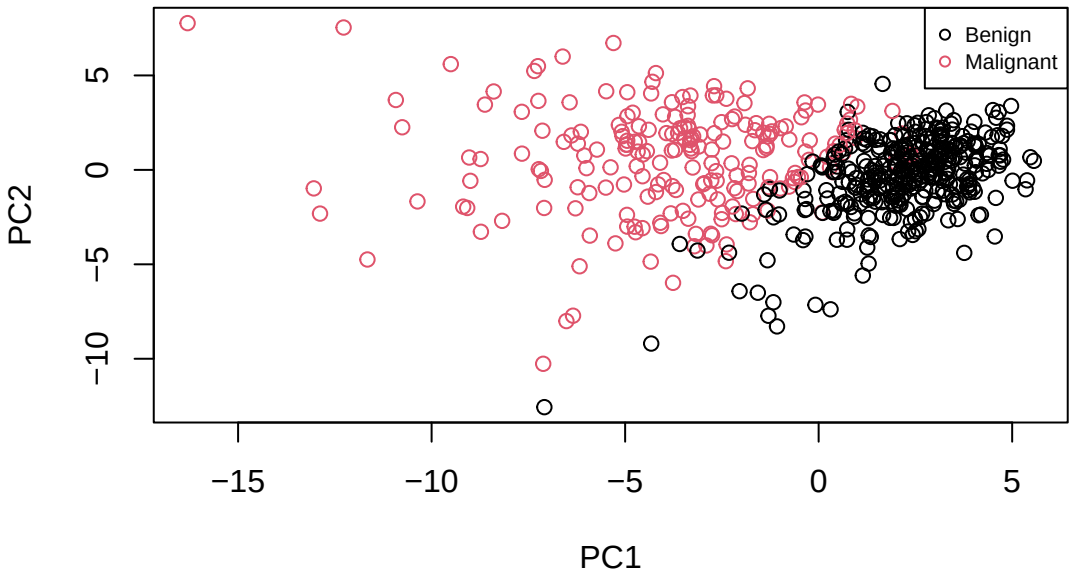
f) What are the coordinates of the first picture (first row) expressed in PC-coordinates? Only give the coordinates of the first three PCs. (**Hint**: `pca_obj$x`)

```
pca_obj$x[1,1:3]
```

```
##       PC1       PC2       PC3
## -9.184755 -1.946870 -1.122179
```

g) Now perform a dimensionality reduction by looking only at the first two PCs. Plot the values of the two first PCs in a scatterplot and make it so the points' colour represents the diagnosis of the pictures. What can we observe in the plot?

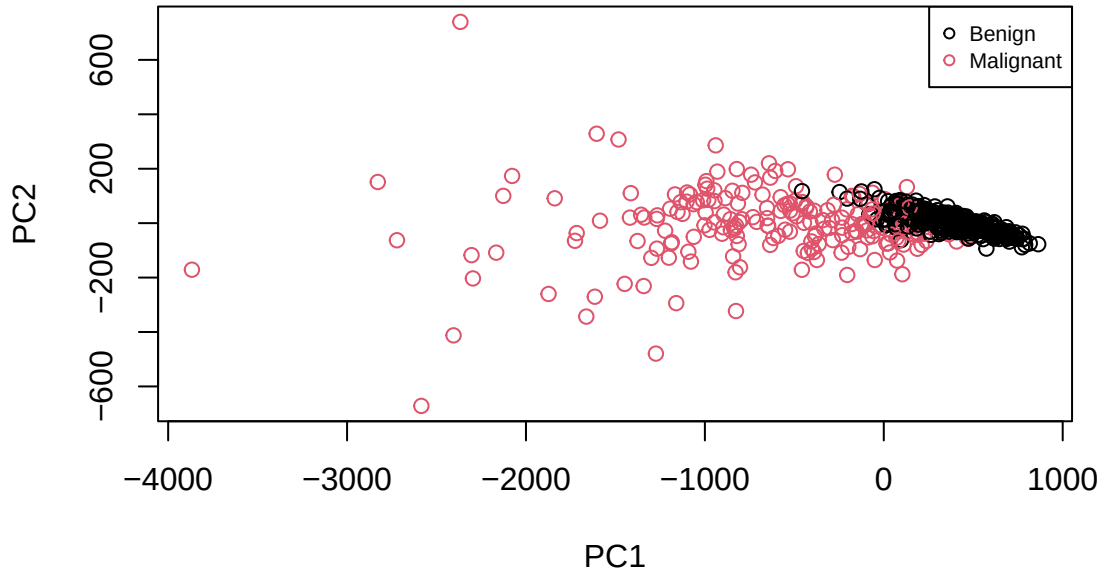
```
plot(PC2~PC1, data=pca_obj$x, col=dat$diagnosis)
legend('topright', c('Benign', 'Malignant'), pch=c(1,1),col=c(1,2), cex=0.7)
```



It seems that in order to differentiate malignant from benign pictures, the 30 variables are not necessarily needed. The two classes are quite well separated even in a two-dimensional representation based on the PCA.

h) We now want to see how the 2D representation changes when no scaling is used. Perform the PCA again but without scaling and create the 2D plot of the first two PCs. What can you observe?

```
pca_noScal <- prcomp(dat[,-c(1,2)], scale. = FALSE)
plot(PC2~PC1, data=pca_noScal$x, col=dat$diagnosis)
legend('topright', c('Benign', 'Malignant'), pch=c(1,1),col=c(1,2), cex=0.7)
```

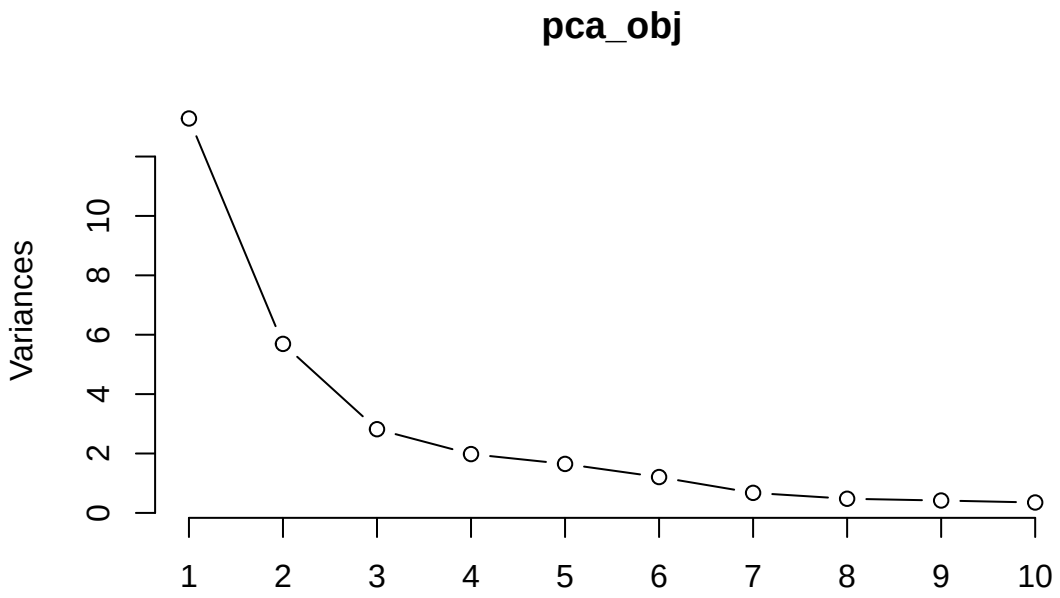


The plot does look a bit different from when performing the PCA with scaling. The malignant data points are much more spread than before. Also, it seems that the benign and malignant pictures are overlapping

more with each other than before.

i) Using the PCs generated with scaling, we wish to find out how many PCs are needed to explain the data well. Create a scree-plot and decide how many PCs should be chosen.

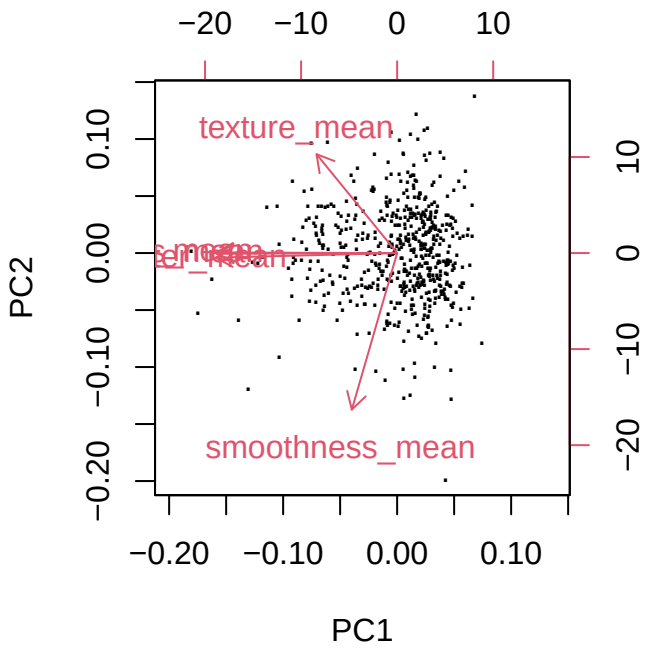
```
screeplot(pca_obj, type = 'l')
```



A possible bend in the curve is occuring at around the 3rd or 4th PC. Thus, one could choose to include the first two or three PCs.

j) In order to make the situation a bit easier to oversee, we perform the PCA again but only with the first five variables (again excluding `id` and `diagnosis`, use scaling). Create a biplot of this PCA, showing the projections of the original variables. Which of the variables has the strongest contribution to PC2?

```
pca_five <- prcomp(dat[,c(3:7)], scale. = TRUE)
biplot(pca_five, xlabs=rep('.', nrow(dat)))
```



Although some labels are a bit difficult to read we can see, that `smoothness_mean` is most stongly contributing to the second PC.