PCA analysis

Andreas Hadjiprocopis

February, 2018

```
#!/usr/bin/env Rscript
```

```
set.seed(1234)

source('lib/UTIL.R');
source('lib/DATA.R');
source('lib/IO.R');
source('lib/TS.R');
source('lib/MC.R');
source('lib/SEASON.R');
source('lib/MIXTURES.R');

infile='cleaned_data/dat1.eliminateNA.csv'

dat1 <- data.frame(read_data(
    filename=infile
))
```

```
## read_data(): data read from file 'cleaned_data/dat1.eliminateNA.csv'.
```

```
if( is.null(dat1) ){
    cat("call to read_data() has failed for file '",infile,"'.\n", sep='')
    quit(status=1)
}
dummy <- remove_columns(inp=dat1, colnames_to_remove=c('id'))
dat1_noid <- dummy[['retained']]
dat1_id <- dummy[['removed']]
dat1_detrended_id <- list(c(dat1_id[['id']][2:length(dat1_id[['id']])]))
names(dat1_detrended_id) <- c('id')
# detrend the data
dat1_detrended <- detrend_dataset(inp=dat1_noid,times=1)
```

PCA In preparing this report I have followed the route of http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

```
library(FactoMineR)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

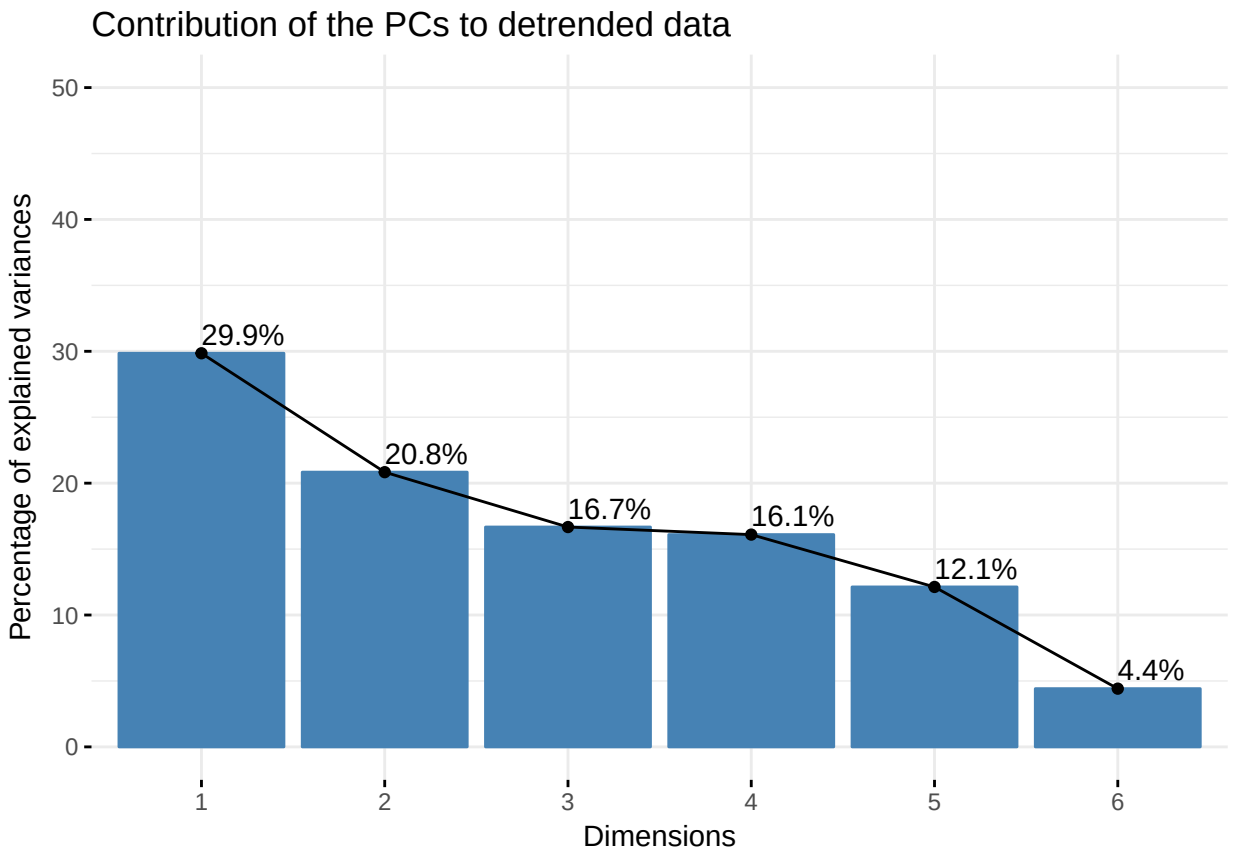
```
adf <- list2dataframe(dat1_detrended)
ncols=ncol(adf)
pcaobj <- PCA(
    adf,
    scale.unit=T, # unit variance, zero mean
    graph=F,
    ncp=ncols
)
```

```
print(get_eigenvalue(pcaobj))
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  1.7910283        29.850472                     29.85047
## Dim.2  1.2500565        20.834275                     50.68475
## Dim.3  1.0003289        16.672148                     67.35690
## Dim.4  0.9659387        16.098978                     83.45587
## Dim.5  0.7277402        12.129003                     95.58488
## Dim.6  0.2649074         4.415123                    100.00000
```

it does not look like there are redundant variables perhaps with the exception of the last component which only contributes 4% to the total variance. If we had to reduce the dimensionality of the original data I would recommend keeping the first 4 components. visualisng the contribution of each component to the variance of the data:

```
fviz_eig(pcaobj, addlabels = TRUE, ylim = c(0, 50), main='Contribution of the PCs to detrended data')
```

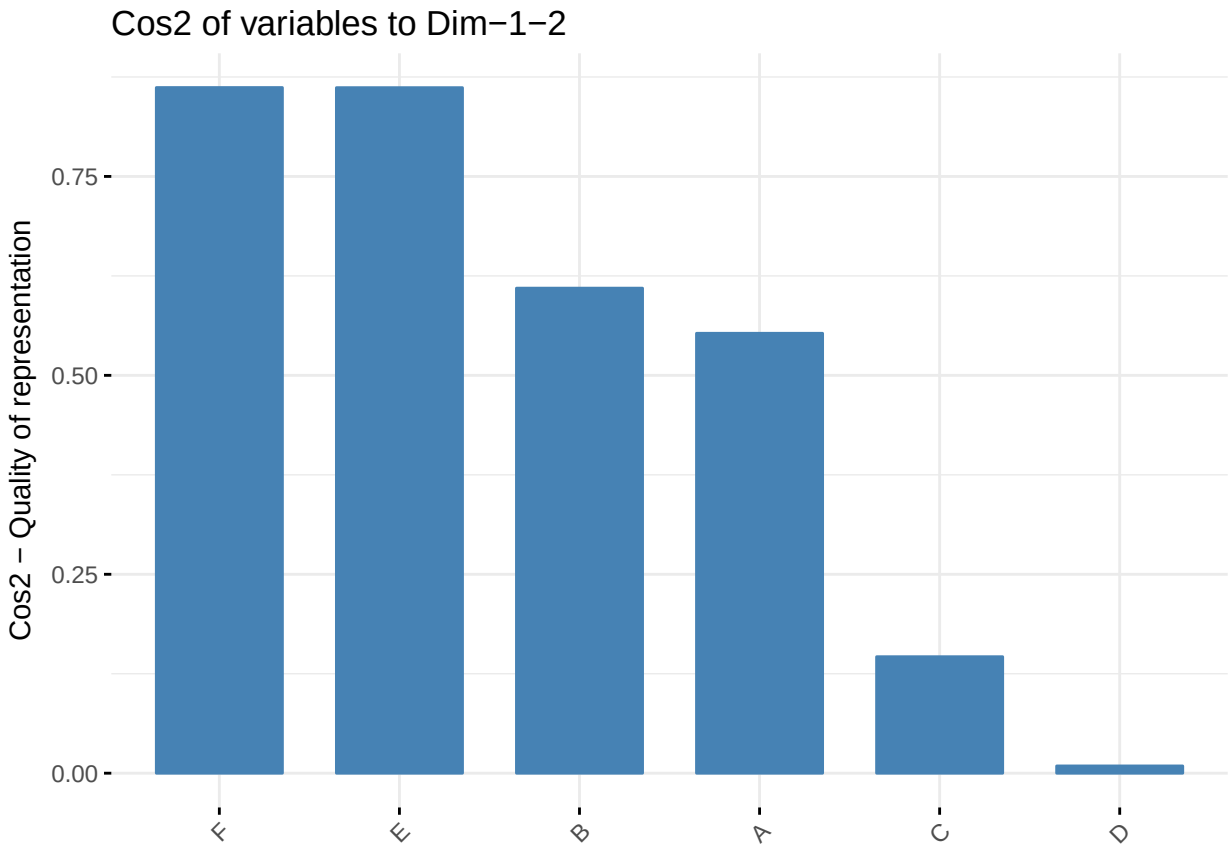


A plot of how is each variable represented by each principal component:

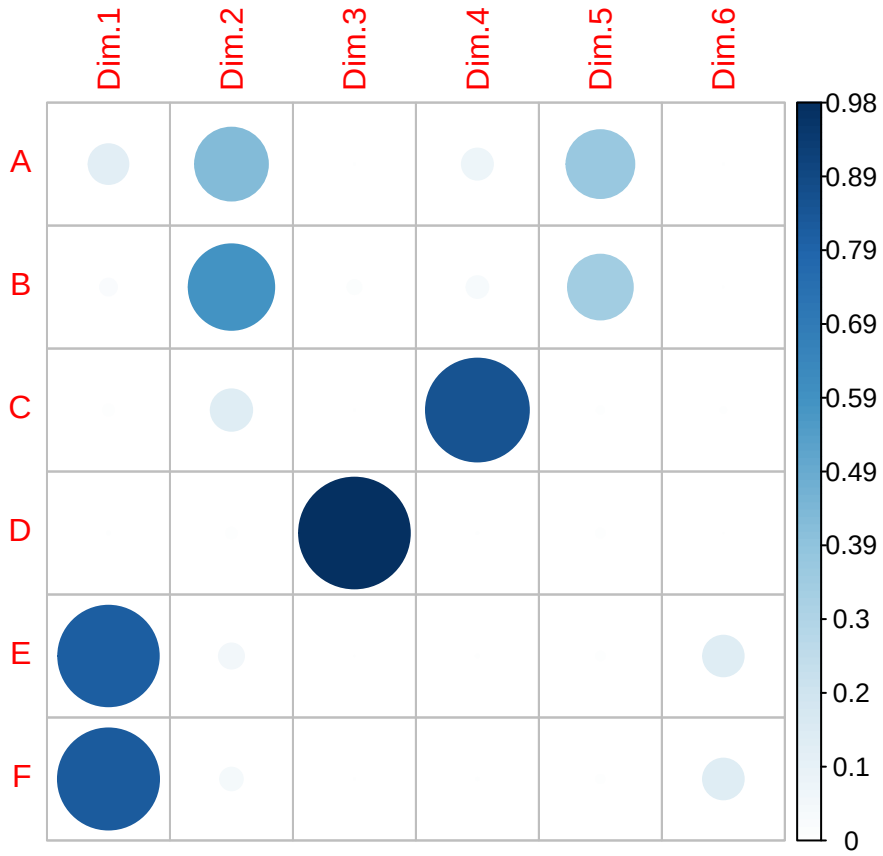
```
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```
fviz_cos2(pcaobj, choice = "var", axes = 1:2)
```

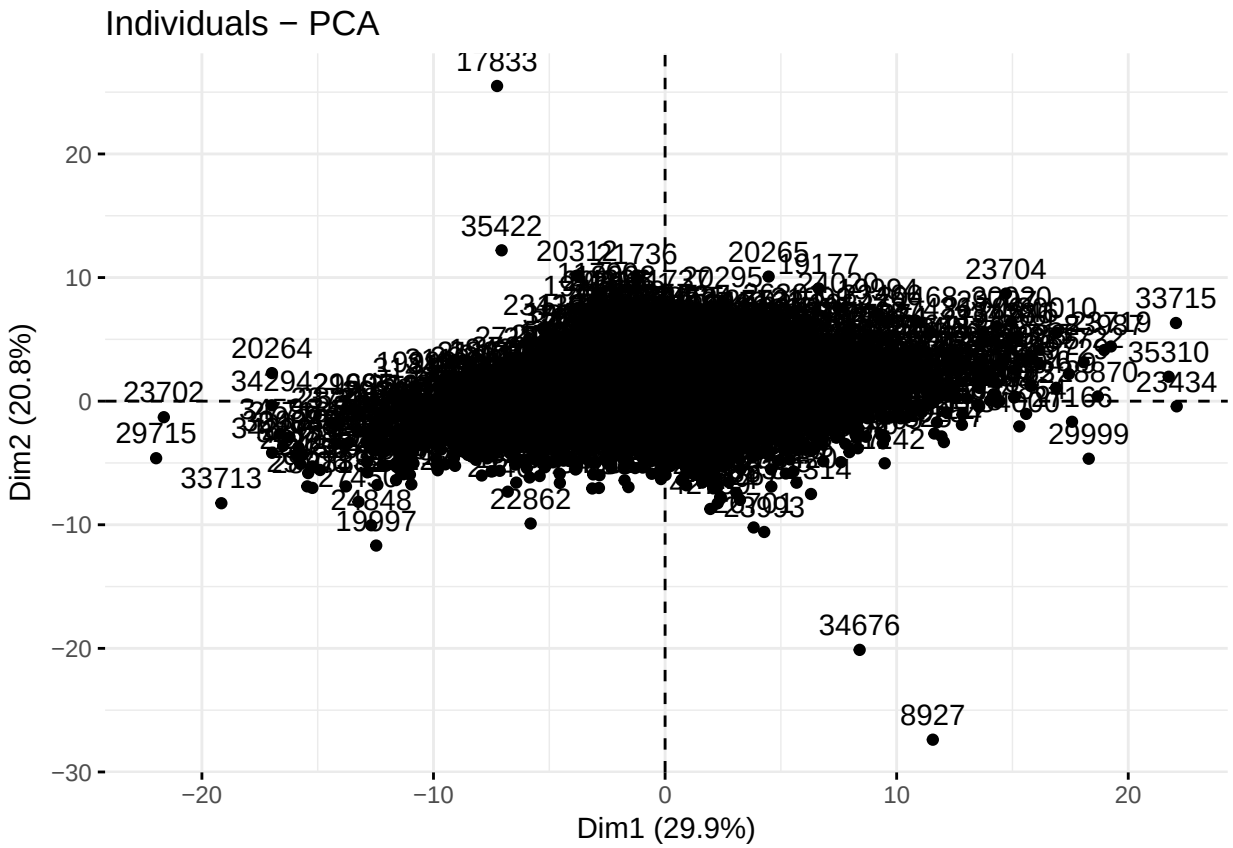



```
var <- get_pca_var(pcaobj)
corrplot(var$cos2, is.corr=FALSE)
```



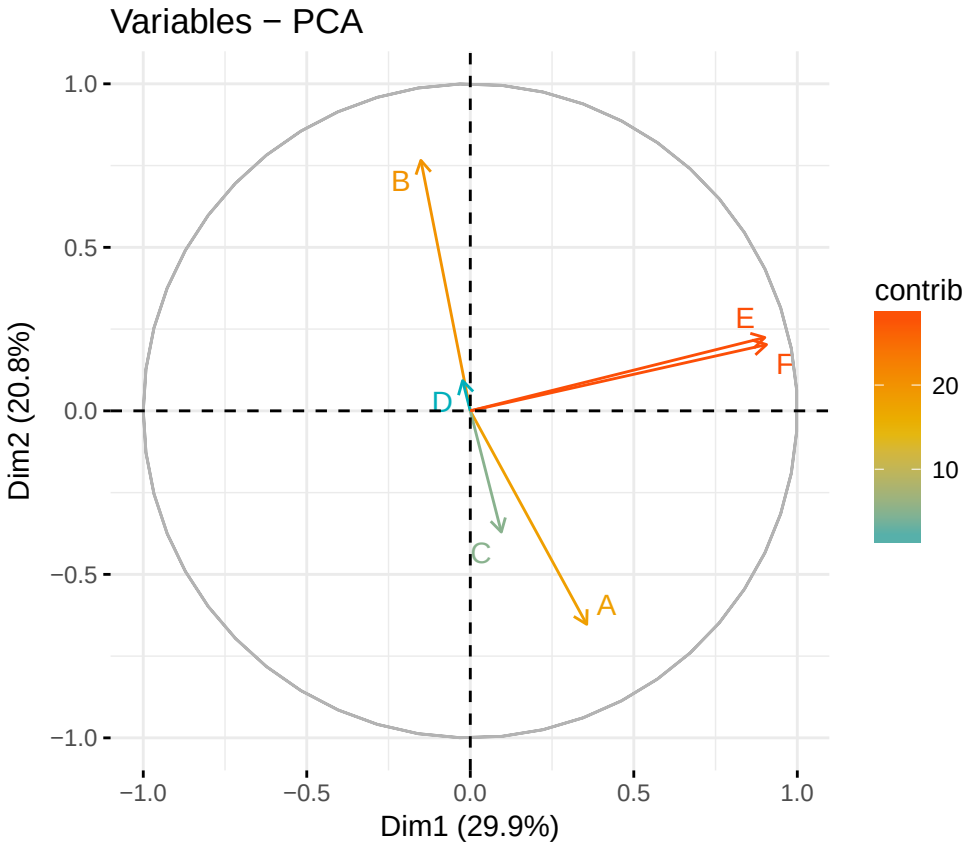
visualisation: a plot of the first 2 components.

```
fviz_pca_ind(pcaobj)
```



variables with shorter vectors are less important for the first components variables with aligned vectors are positively correlated.

```
fviz_pca_var(
    pcaobj, col.var = "contrib",
    gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
    repel = TRUE
)
```

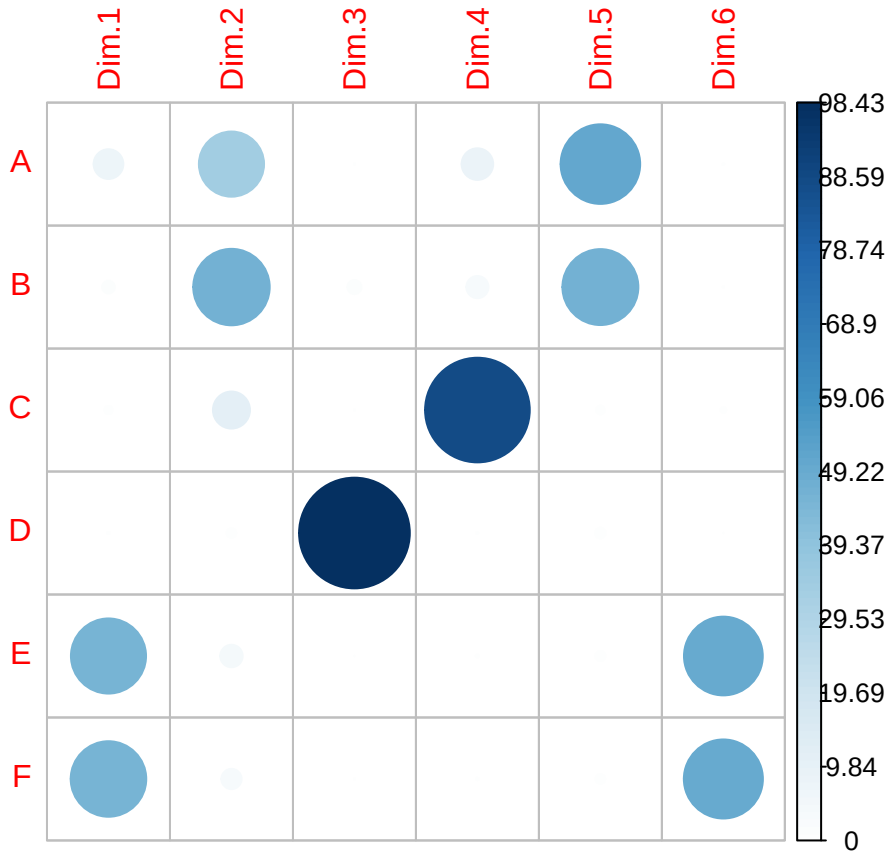


contribution of the variables to the variance of the data (i.e. importance of the 'real' variables):

```
print(var$contrib)
```

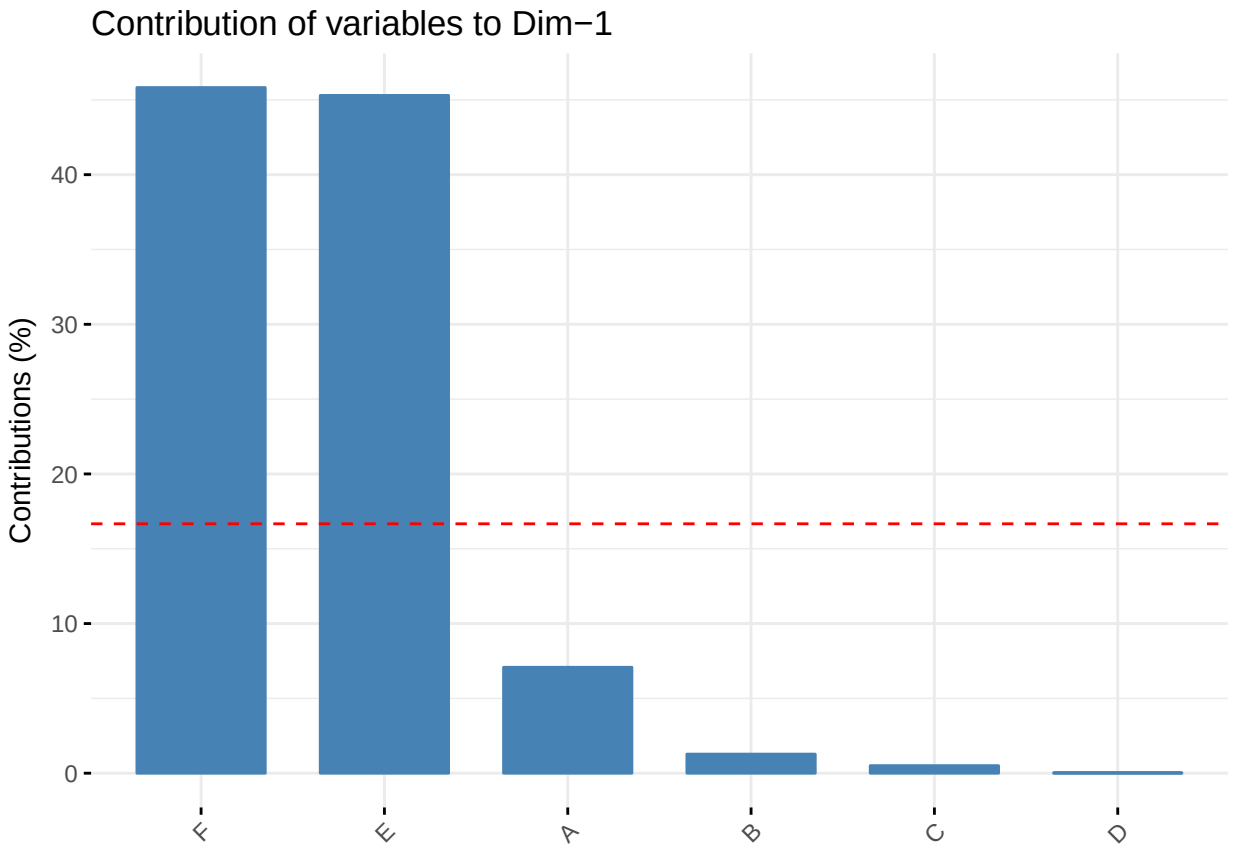
```
##         Dim.1      Dim.2        Dim.3       Dim.4      Dim.5        Dim.6
## A  7.07733359 34.0843255 6.375678e-03  7.99481606 50.8195298 1.761936e-02
## B  1.27466851 46.9385566 1.554029e+00  3.91886295 46.3069450 6.938345e-03
## C  0.50277472 10.9883708 6.305069e-03 87.95297911  0.5495663 3.945954e-06
## D  0.03097688  0.6833524 9.843002e+01  0.02230487  0.8331156 2.323091e-04
## E 45.29450517  4.0236214 3.761742e-04  0.07399062  0.7958687 4.981164e+01
## F 45.81974114  3.2817734 2.896544e-03  0.03704639  0.6949745 5.016357e+01
```

```
corrplot(var$contrib, is.corr=FALSE)
```



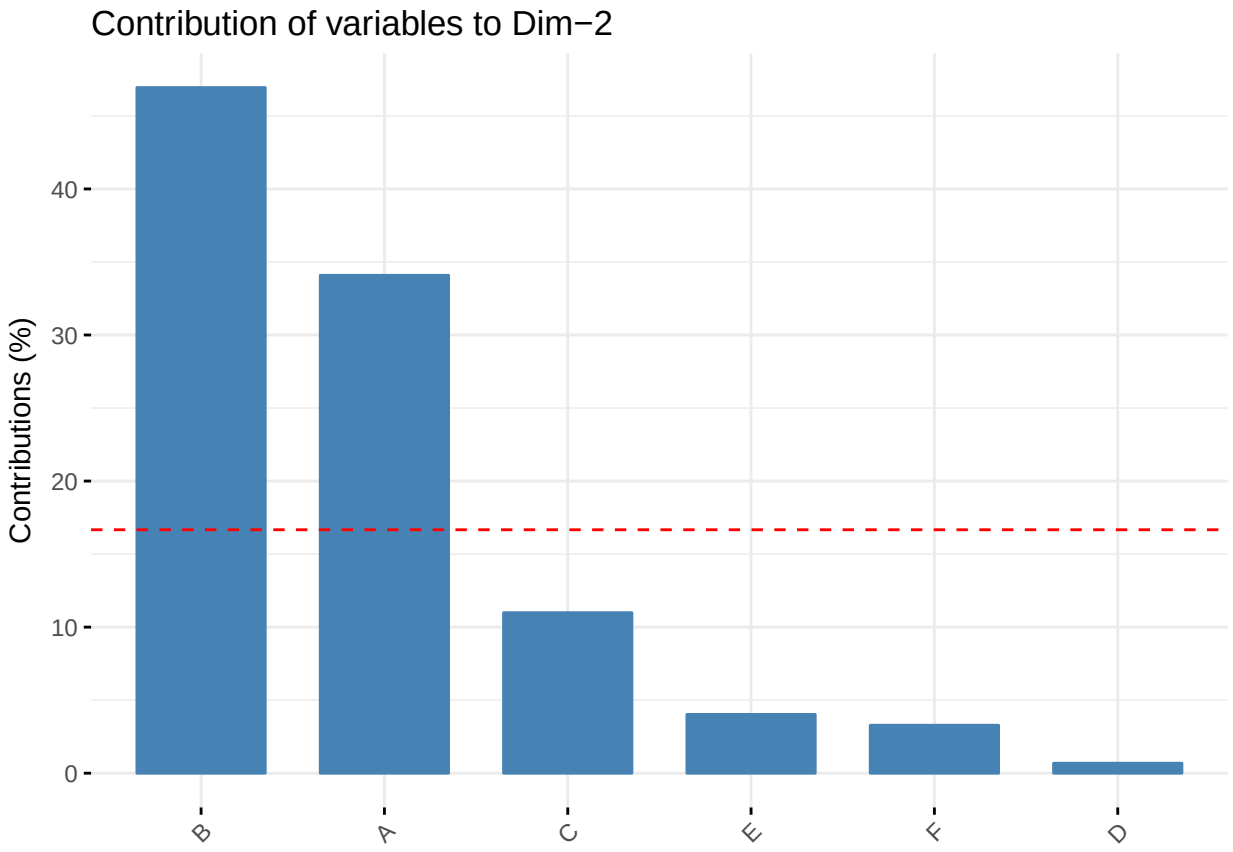
Contributions of variables to PC1

```
fviz_contrib(pcaobj, choice = "var", axes = 1, top = 10)
```



Contributions of variables to PC2

```
fviz_contrib(pcaobj, choice = "var", axes = 2, top = 10)
```


Contribution of variables to Dim−2
Contributions (%)
40
30
20
10
0
B
A
C
E
F
D