Kurs K1 (Wprowadzenie do eksploracji danych) – Laboratorium 7

Redukcja wymiarowości danych metodą PCA

Przykład: Przeprowadzenie analizy składowych głównych dla zbioru „iris".

Instalujemy potrzebne pakiety i importujemy biblioteki:

```
> install.packages(c("FactoMineR", "factoextra", "corrplot"))
> library("FactoMineR")       ## do PCA
> library("factoextra")       ## do PCA
> library("corrplot")         ## do zilustrowania macierzy korelacji na wykresie
```

Wczytujemy dane ze zbioru "iris":

```
> library(datasets)
> data(iris)                          ## wczytanie danych
> colnames(iris)<-c("dl_dzialki","szer_dzialki","dl_platka","szer_platka","gatunek")
                              ## polskie nazwy cech
> View(iris)
```

Standaryzujemy zmienne (bierzemy tylko pierwsze cztery kolumny, z wartościami liczbowymi):

```
> iris.scaled <- scale(iris[,1:4], center = TRUE, scale = TRUE)
> View(iris.scaled)
```

Wyznaczamy macierz korelacji:

```
> res.cor <- cor(iris.scaled)
> View(res.cor)
> corrplot(res.cor)                   ## wykres
```

```
> res.cor
             dl_dzialki szer_dzialki  dl_platka szer_platka
dl_dzialki    1.0000000   -0.1175698  0.8717538   0.8179411
szer_dzialki -0.1175698    1.0000000 -0.4284401  -0.3661259
dl_platka     0.8717538   -0.4284401  1.0000000   0.9628654
szer_platka   0.8179411   -0.3661259  0.9628654   1.0000000
```

Przeprowadzenie analizy składowych głównych (PCA):

```
> res.pca <- PCA(iris.scaled, graph = TRUE, ncp = 4)
                              ## ncp to liczba trzymanych wymiarów
> print(res.pca)              ## Wyniki PCA są dostępne w zmiennej res.pca
```

```
> print(res.pca)
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 150 individuals, described by 4 variables
*The results are available in the following objects:

   name               description
1  "$eig"             "eigenvalues"
2  "$var"             "results for the variables"
3  "$var$coord"       "coord. for the variables"
4  "$var$cor"         "correlations variables - dimensions"
5  "$var$cos2"        "cos2 for the variables"
6  "$var$contrib"     "contributions of the variables"
7  "$ind"             "results for the individuals"
8  "$ind$coord"       "coord. for the individuals"
9  "$ind$cos2"        "cos2 for the individuals"
10 "$ind$contrib"     "contributions of the individuals"
11 "$call"            "summary statistics"
12 "$call$centre"     "mean of the variables"
13 "$call$ecart.type" "standard error of the variables"
14 "$call$row.w"      "weights for the individuals"
15 "$call$col.w"      "weights for the variables"
```

Wyniki:

```
> var <- get_pca_var(res.pca)
> var

Principal Component Analysis Results for variables
 ===================================================
  Name       Description
1 "$coord"   "Coordinates for the variables"
2 "$cor"     "Correlations between variables and dimensions"
3 "$cos2"    "Cos2 for the variables"
4 "$contrib" "contributions of the variables"
```

```
> res.pca$eig
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 2.91849782             72.9624454                            72.96245
comp 2 0.91403047             22.8507618                            95.81321
comp 3 0.14675688              3.6689219                            99.48213
comp 4 0.02071484              0.5178709                           100.00000

> res.pca$var
$coord
                  Dim.1      Dim.2       Dim.3       Dim.4
dl_dzialki    0.8901688 0.36082989 -0.27565767 -0.03760602
szer_dzialki -0.4601427 0.88271627  0.09361987  0.01777631
dl_platka     0.9915552 0.02341519  0.05444699  0.11534978
szer_platka   0.9649790 0.06399985  0.24298265 -0.07535950

$cor
                  Dim.1      Dim.2       Dim.3       Dim.4
dl_dzialki    0.8901688 0.36082989 -0.27565767 -0.03760602
szer_dzialki -0.4601427 0.88271627  0.09361987  0.01777631
dl_platka     0.9915552 0.02341519  0.05444699  0.11534978
szer_platka   0.9649790 0.06399985  0.24298265 -0.07535950

$cos2
                 Dim.1       Dim.2       Dim.3        Dim.4
dl_dzialki   0.7924004 0.130198208 0.075987149 0.0014142127
szer_dzialki 0.2117313 0.779188012 0.008764681 0.0003159971
dl_platka    0.9831817 0.000548271 0.002964475 0.0133055723
szer_platka  0.9311844 0.004095980 0.059040571 0.0056790544

$contrib (wkłady zmiennych)
                 Dim.1       Dim.2     Dim.3     Dim.4
dl_dzialki   27.150969 14.24440565 51.777574  6.827052
szer_dzialki  7.254804 85.24748749  5.972245  1.525463
dl_platka    33.687936  0.05998389  2.019990 64.232089
szer_platka  31.906291  0.44812296 40.230191 27.415396
```

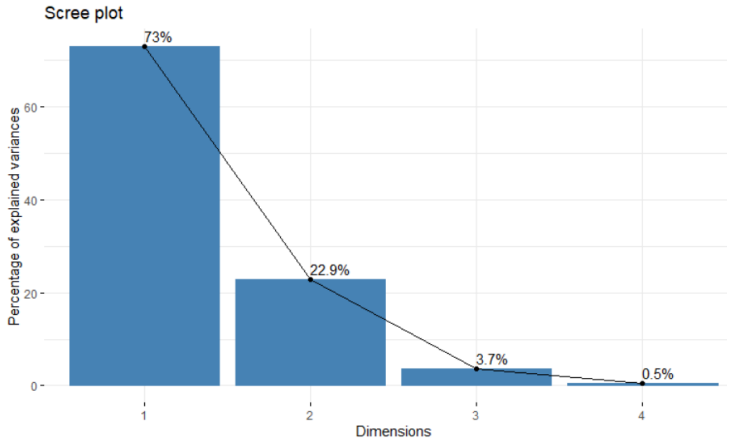
Wartości własne, % wyjaśnianej wariancji:

```
> eig.val<-get_eigenvalue(res.pca)
> colnames(eig.val)<-c("wartosc_wlasna","proc_wariancji","proc_skum_wariancji")
> View(eig.val)
```

```
> eig.val
      wartosc_wlasna proc_wariancji proc_skum_wariancji
Dim.1     2.91849782     72.9624454            72.96245
Dim.2     0.91403047     22.8507618            95.81321
Dim.3     0.14675688      3.6689219            99.48213
Dim.4     0.02071484      0.5178709           100.00000
```

Wykres osypiskowy:

```
> fviz_eig(res.pca, ncp = 4, addlabels = TRUE)
```



Ładunki czynnikowe (korelacje między zmiennymi pierwotnymi i składowymi głównymi):

```
> ladunki_czynnikowe<-res.pca$var$cor
> View(ladunki_czynnikowe)

> fviz_pca_var(res.pca, col.var = "black") ## wykres ładunków dla dwóch
                                             pierwszych składowych
> corrplot(res.pca$var$cor, is.corr=FALSE) ## wykres korelacji dla wszystkich
```

```
> ladunki_czynnikowe
                 Dim.1      Dim.2       Dim.3       Dim.4
dl_dzialki     0.8901688 0.36082989 -0.27565767 -0.03760602
szer_dzialki -0.4601427 0.88271627  0.09361987  0.01777631
dl_platka      0.9915552 0.02341519  0.05444699  0.11534978
szer_platka    0.9649790 0.06399985  0.24298265 -0.07535950
```

Zasoby zmienności wspólnej:

```
> zasoby<-res.pca$var$cos2*100
> for(i in 2:4)
   zasoby[,i]<-zasoby[,i]+zasoby[,i-1]
> View(zasoby)
```

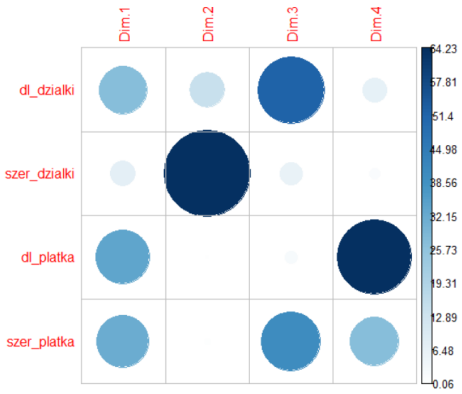
```
> zasoby
                Dim.1    Dim.2    Dim.3 Dim.4
dl_dzialki    79.24004 92.25986 99.85858   100
szer_dzialki 21.17313 99.09193 99.96840   100
dl_platka     98.31817 98.37300 98.66944   100
szer_platka   93.11844 93.52804 99.43209   100
```

Wkłady zmiennych w składowe główne [%]:

```
> wklady_zm<-res.pca$var$contrib
> colnames(wklady_zm)<-c("Dim.1","Dim.2","Dim.3","Dim.4")
> View(wklady_zm)
```

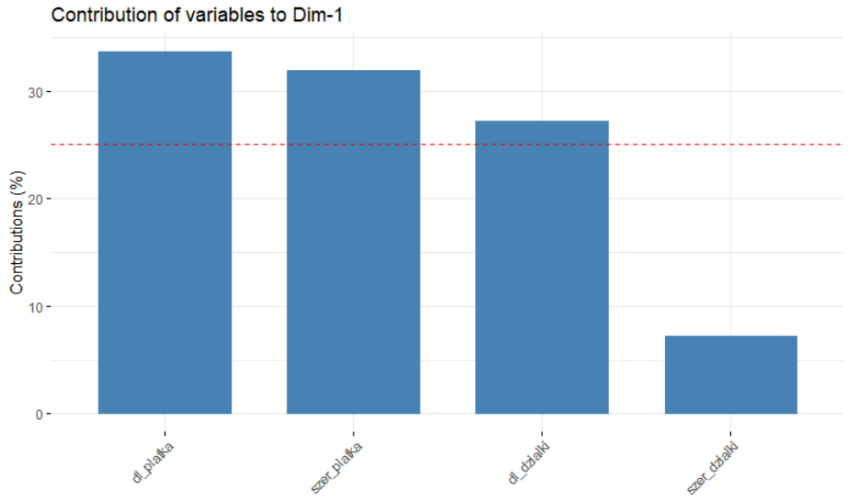
```
> wklady_zm
                 Dim.1       Dim.2     Dim.3     Dim.4
dl_dzialki    27.150969 14.24440565 51.777574  6.827052
szer_dzialki  7.254804 85.24748749  5.972245  1.525463
dl_platka     33.687936  0.05998389  2.019990 64.232089
szer_platka   31.906291  0.44812296 40.230191 27.415396
```

```
> corrplot(res.pca$var$contrib, is.corr=FALSE)  ## wykres
```

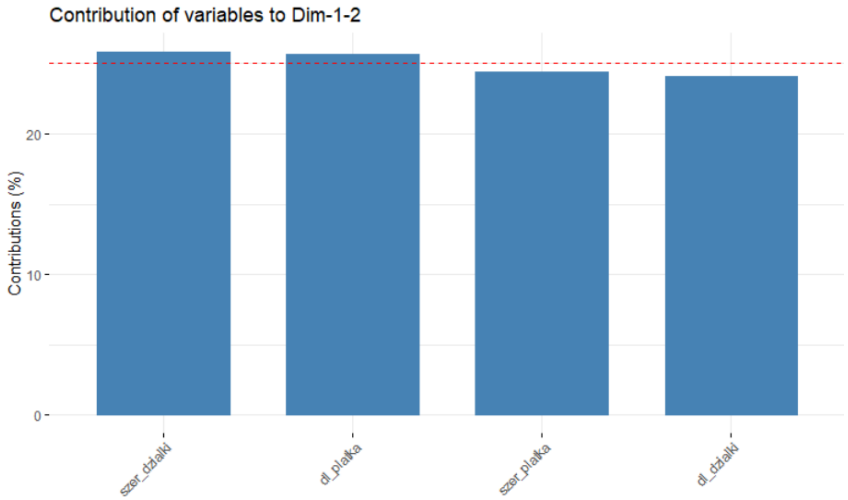


Wykresy łącznego wkładu zmiennych do głównych składowych (dla liczby wymiarów od 1 do czterech):

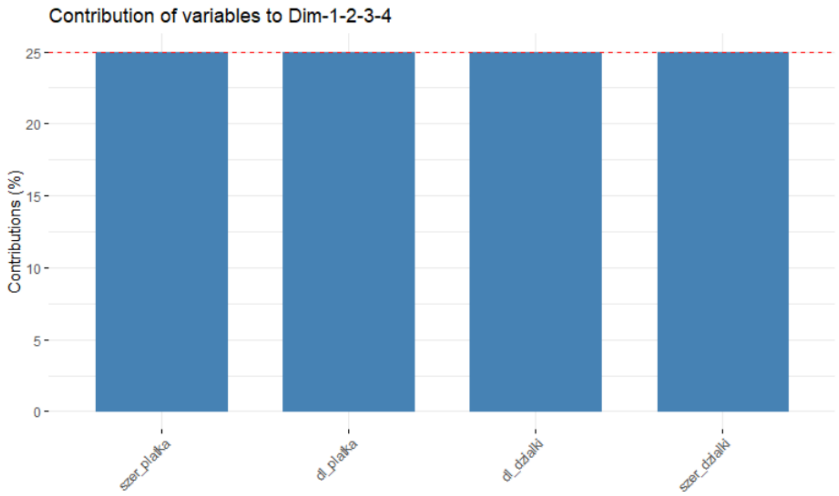
```
## top to maksymalna liczba cech pokazanych na wykresie
> fviz_contrib(res.pca, choice = "var", axes = 1, top = 4)      ## dla pierwszej
                                                                ##składowej głównej (1
                                                                wymiaru)
```




```
> fviz_contrib(res.pca, choice = "var", axes = 1:2, top = 4)          ## dla
                                                              dwóch ##pierwszych składowych (2
                                                              wymiarów)
```



```
> fviz_contrib(res.pca, choice = "var", axes = 1:4, top = 4)     ## dla czterech
                                                                  ## pierwszych składowych
```



Wyznaczamy **wektory własne**:

```
> res.eig <- eigen(res.cor)
> wektory_wlasne<-res.eig$vectors
> colnames(wektory_wlasne)<-c("Dim.1","Dim.2","Dim.3","Dim.4")
> rownames(wektory_wlasne)<-c("dl_dzialki","szer_dzialki","dl_platka","szer_platka")
> View(wektory_wlasne)
```

```
> wektory_wlasne
                 Dim.1       Dim.2      Dim.3      Dim.4
dl_dzialki    0.5210659 -0.37741762  0.7195664  0.2612863
szer_dzialki -0.2693474 -0.92329566 -0.2443818 -0.1235096
dl_platka     0.5804131 -0.02449161 -0.1421264 -0.8014492
szer_platka   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Albo tak:

```
> res.eig <- eigen(res.cor)
> res.eig
eigen() decomposition
$values
[1] 2.91912926 0.91184362 0.14426500 0.02476212

$vectors
           [,1]        [,2]       [,3]       [,4]
[1,]  0.5210309 -0.37921152  0.7198899  0.2578448
[2,] -0.2713291 -0.92251432 -0.2458120 -0.1221652
[3,]  0.5795399 -0.02547068 -0.1458335 -0.8013847
[4,]  0.5648371 -0.06721014 -0.6325089  0.5257132
```

Tworzymy **nową tabelę danych** ("iris_pca") na podstawie wyników PCA (bierzemy tylko dwie pierwsze składowe główne jako nowe cechy i doklejamy kolumnę ze zmienną celu (gatunek irysa):

```
> iris.pca<-res.pca$ind$coord[,-(3:4)]     ## kopiujemy dwie pierwsze nowe cechy
> iris.pca<-cbind(iris.pca, as.character(iris[,5]))  ## dodajemy kolumnę z klasą
                                    ##(kolumna nr 5 "gatunek" z tabeli "iris")
## można też tak:cbind(iris.pca, as.character(iris$gatunek))
> colnames(iris.pca)<-c("PC1","PC2","gatunek")
> View(iris.pca)
```

```
> iris.pca
     PC1                   PC2                    gatunek
1   "-2.26470280880758"   "0.480026596520989"    "setosa"
2   "-2.08096115196577"   "-0.674133556605352"   "setosa"
3   "-2.3642290538903"    "-0.341908023884675"   "setosa"
```

```
4   "-2.29938421704271"   "-0.597394507674675"  "setosa"
```

Wyniki PCA dla poszczególnych (indywidualnych) obserwacj (wierszy tabeli *iris*)::

```
> res.pca$ind
$coord
          Dim.1        Dim.2        Dim.3        Dim.4
1   -2.26470281  0.480026597 -0.127706022 -0.024168204
2   -2.08096115 -0.674133557 -0.234608854 -0.103006775
3   -2.36422905 -0.341908024  0.044201485 -0.028377053
4   -2.29938422 -0.597394508  0.091290106  0.065955560
…

$cos2
            Dim.1        Dim.2        Dim.3        Dim.4
1   9.539975e-01 4.286032e-02 3.033525e-03 1.086460e-04
2   8.927725e-01 9.369248e-02 1.134754e-02 2.187482e-03
3   9.790410e-01 2.047578e-02 3.422122e-04 1.410446e-04
4   9.346682e-01 6.308947e-02 1.473268e-03 7.690193e-04
…

$contrib
            Dim.1        Dim.2        Dim.3        Dim.4
1   1.171580e+00 1.680655e-01 0.0740854699 0.0187981878
2   9.891845e-01 3.314667e-01 0.2500340065 0.3414749194
3   1.276816e+00 8.526419e-02 0.0088753196 0.0259156335
4   1.207737e+00 2.602978e-01 0.0378580037 0.1400006497
…
```

Zadanie do wykonania

1. Sprawdź na zbiorze „iris” działanie kodu R przedstawionego w niniejszym opracowaniu.
2. Następnie wykonaj podobne instrukcje dla zbioru danych *wine* (dla zbioru bez obserwacji brakujących i bez punktów oddalonych). Odpowiedz na poniższe pytania/polecenia. W pliku przygotuj odpowiednie informacje (wklej polecenie R oraz wyniki/rysunki):
   a) Wyświetl wartości własne składowych głównych wraz z procentem wyjaśnianej wariancji i procentem skumulowanej wariancji oraz wykres osypiskowy.
   b) Wyświetl ładunki czynnikowe.
   c) Wyświetl zasoby zmienności wspólnej.
   d) Wyświetl wkłady zmiennych pierwotnych w poszczególne składowe główne. Zilustruj je na wykresie korelacji.
   e) Sporządź wykres łącznego wkładu zmiennych do głównych składowych dla pierwszych trzech wymiarów (Dim-1-2-3).
   f) Wyświetl wektory własne.
   g) Ile składowych głównych należy pozostawić w nowym modelu? Uzasadnij wykorzystując: procent wyjaśnionej wariancji, kryterium Kaisera, wykres osypiska i zasoby zmienności wspólnej (przedstaw wyniki dla wszystkich czterech kryteriów i ostateczny wybór).
   h) Utwórz nową tabelę danych z odpowiednią liczbą nowych zmiennych/cech (składowych głównych) oraz ze zmienną celu (klasy). Załącz wynik polecenia View. Zapisz wygenerowany zbiór do pliku (przyślij ten plik mailem).