# PCA - Analiza składowych głównych

Patryk Nizio

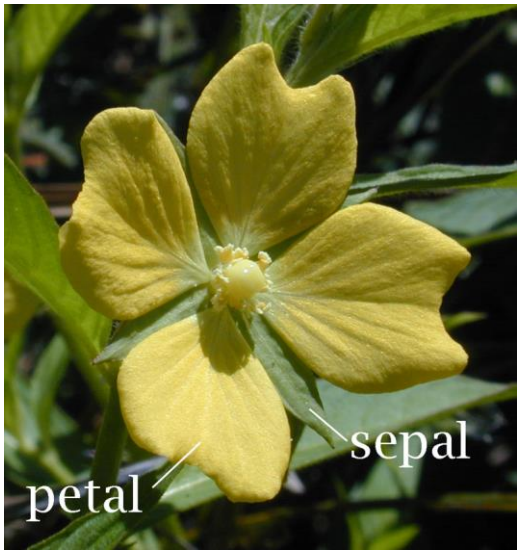
7 maja 2019

## PCA - wstęp

Analiza składowych głównych (PCA) służy m.in. do redukcji liczby zmiennych opisujących zjawiska, czy do odkrycia prawidłowości między zmiennymi. Polega ona na wyznaczeniu składowych będących kombinacją liniową badanych zmiennych. Dokładna analiza składowych głównych umożliwia wskazanie tych zmiennych początkowych, które mają duży wpływ na wygląd poszczególnych składowych głównych czyli tych, które tworzą grupę jednorodną. Składowa główna (u której wariancja jest zmaksymalizowana) jest wówczas reprezentantem tej grupy.

## Zestaw danych - Iris

Zestaw danych "Iris" składa się z 50 próbek od każdego z trzech gatunków irysów (Iris setosa, Iris virginica i Iris versicolor). Z każdej próbki zmierzono cztery cechy: długość i szerokość działek i płatków w centymetrach.




*Flower with petal & sepal*

Przykładowe dane:

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
```

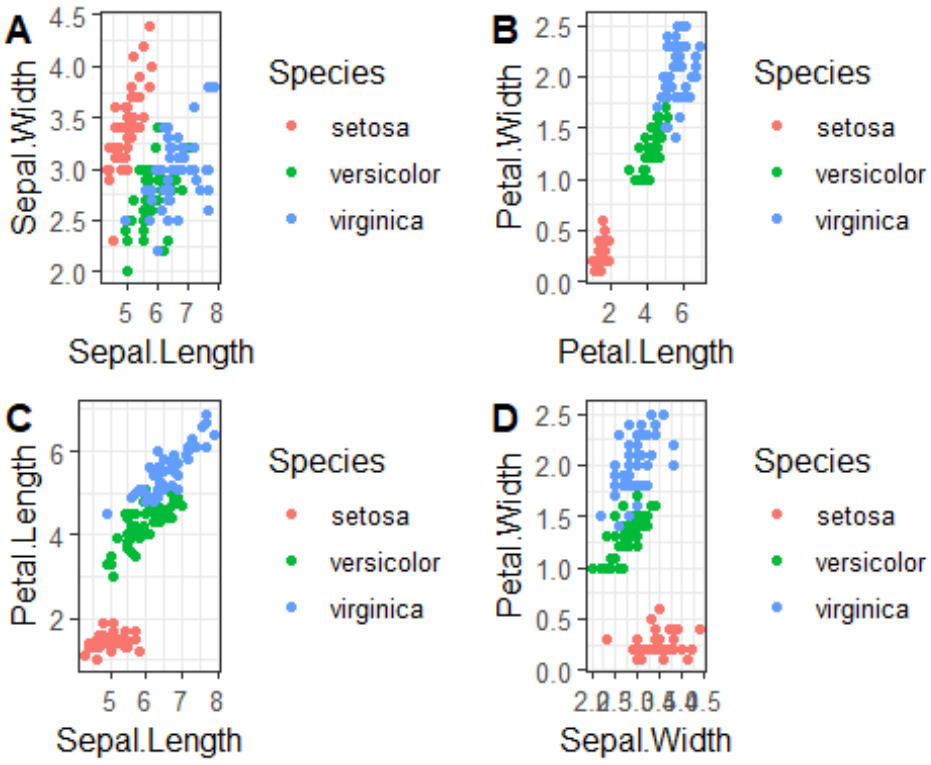
```
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Statystyki opisowe

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

## Wizualizacja danych

Wykresy rozrzutu grupując po odmianie irysa:



## PCA - analiza

### Korelacja

Zbadano korelację metodą pearsona, istnieje silna korelacja między _**Petal.Length** a **Petal.Width** oraz **Sepal.Length** i __Petal.Length__, **Petal.Width**.

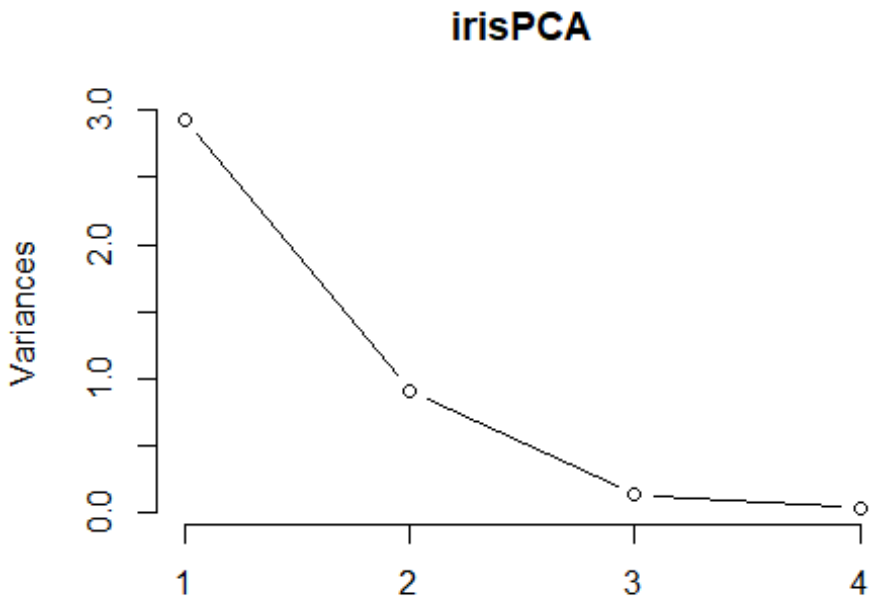
```
cor(iris[1:4], method = "pearson")
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

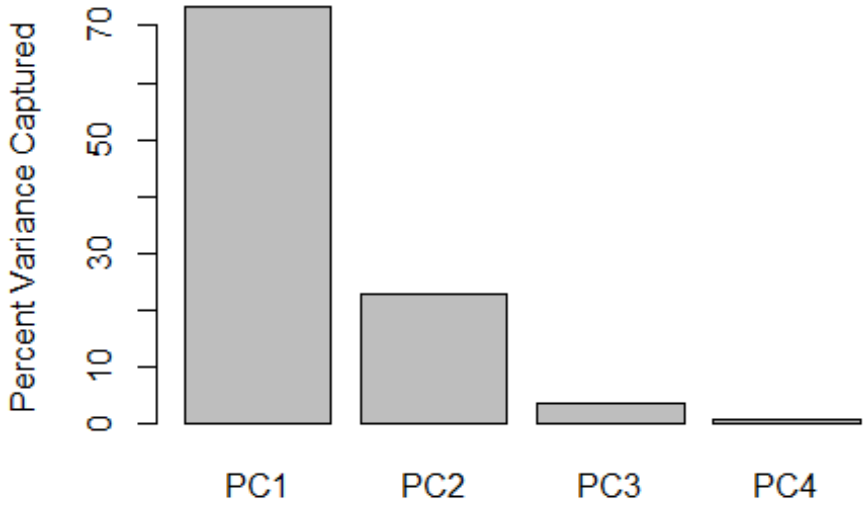
### Wykres osypiska

Jest to wykres liniowy wartości własnych. Po drugim czynniku następuje łagodny spadek, uwzględniamy najbardziej znaczące czynniki czyli pierwszy i drugi.

```
species <- iris[, 5]
logIris <- log(iris[, 1:4])
irisPCA <- prcomp(logIris, center = TRUE, scale. = TRUE)
plot(irisPCA, type = "l")
```



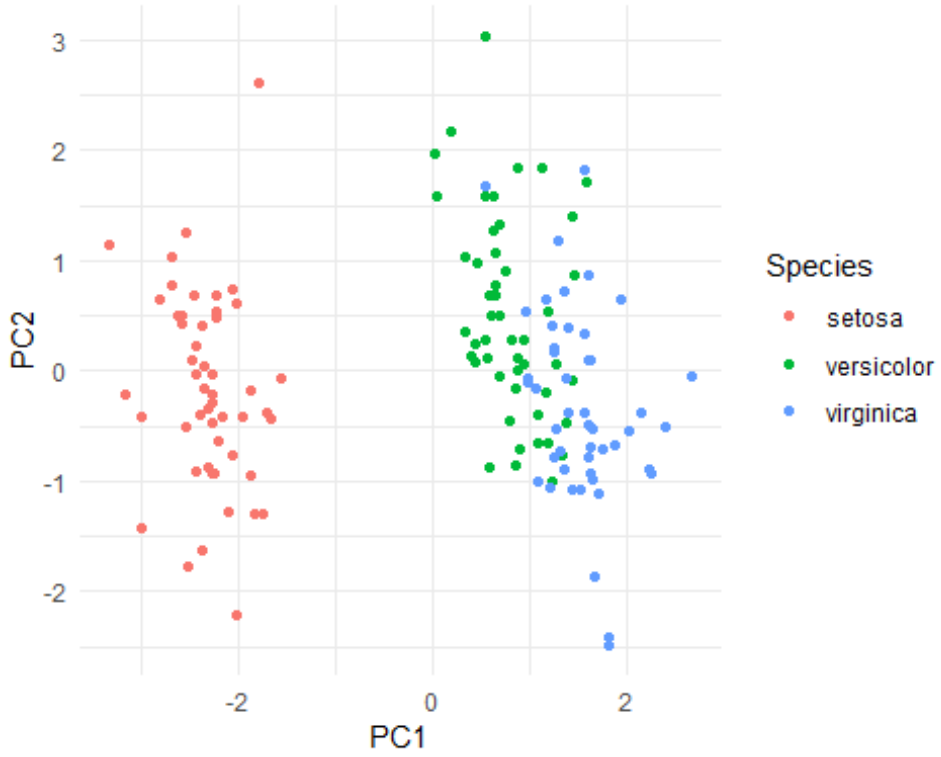

Czynnik pierwszy jak widać na wykresie jest najbardziej znaczący. Natomiast pierwsze dwa czynniki odpowiadają za więcej niż 90% wariancji danych.




**Wyznaczenie wektorów własnych**

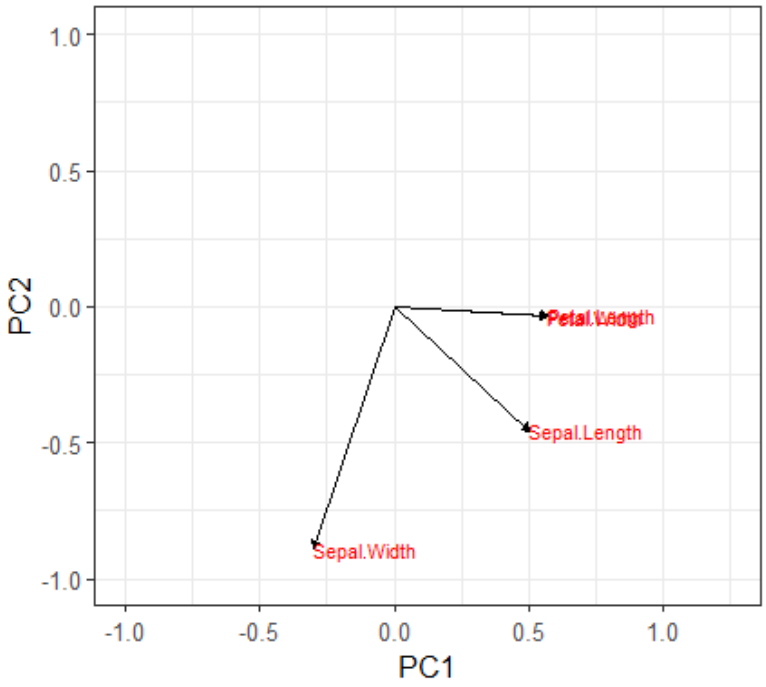
```
## Standard deviations (1, .., p=4):
## [1] 1.7124583 0.9523797 0.3647029 0.1656840
##
## Rotation (n x k) = (4 x 4):
##                     PC1         PC2        PC3         PC4
## Sepal.Length  0.5038236 -0.45499872  0.7088547  0.19147575
## Sepal.Width  -0.3023682 -0.88914419 -0.3311628 -0.09125405
## Petal.Length  0.5767881 -0.03378802 -0.2192793 -0.78618732
## Petal.Width   0.5674952 -0.03545628 -0.5829003  0.58044745
```

**Wykres rozrzutu dla pierwszego i drugiego czynnika**




Poniższy wykres rozrzutu pokazuje nam, ile każda zmienna przyczynia się do każdego głównego czynnika. Na przykład Sepal.Width ma niewielki wpływ na PC1, ale stanowi dużą część PC2, odwrotną sytuację mamy dla Petal.Width który wpływa w małym stopniu na PC2 ale ma duży wpływ na PC1.

**Projekcja zmiennych na płaszczyznę czynników**




## Podsumowanie

Podczas analizy za pomocą PCA mamy możliwość analizowania czynników istotnych dla zjawiska. Możemy zredukować liczbę wymiarów za pomocą niezależnych składowych głównych.

W pierwszym czynniku zmienne **Sepal.Length**, **Petal.Length** i **Petal.Width** są ze sobą skorelowane oraz odgrywają najważnieszą role przy rozpoznaniu odmiany irysów. W drugim czynniku najważnieszy była zmienna_ **Sepal.Width** która pozwała odróżnić odmianę versicolor od virginica.