
**Politechnika Gdańska**
**Wydział Fizyki Technicznej i Matematyki Stosowanej**

# Projekt 1
# Analiza składowych głównych

*Kamil Łangowski*

prowadzący:
dr inż. Anna Szafrańska

21 czerwca 2022

# Spis treści

# 1 Część pierwsza

## Opis problemu

W tej części projektu wykonujemy analizę PCA dla rzeczywistych danych dot. nowotworu gruczołu sutkowego, posługując się napisanymi przez siebie funkcjami (w oparciu o działania na wektorach i macierzach) w środowisku R. Nie skorzystamy z żadnych dedykowanych pakietów do analizy PCA.

## 1.1 Zadanie I.1 (Wczytanie danych.)

### Polecenie

Wczytać dane z pliku *danePCA.csv*. Sprawdzić poprawność wczytanych wartości pod kątem liczby obserwacji, liczby zmiennych i typów zmiennych (polecenia *head()*, *dim()*, *str()*).

W dalszej części mówiąc o danych mamy na myśli wartości zmiennych 3-32 opisujących charakterystyki komórek.

Zapisać jako zmienną **X** macierz danych o wymiarze (569×30).

### Realizacja

Dane wczytano do programu. Ustalono, że zbiór składa się 569 obserwacji i 32 cech. Usunięto pierwsze dwie zmienne: *id* typu całkowitoliczbowego oraz *diagnosis* typu znakowego. Dane oznaczono jako **X**.

## 1.2 Zadanie I.2 (Przygotowanie danych i statystyki opisowe.)

### Polecenie

Wyznaczyć wektor wartości średnich oraz macierz kowariancji dla macierzy danych $X$.

Wystandaryzować dane (centrowanie i skalowanie). Zapisać wystandaryzowane dane jako zmienną $Y$. Wyznaczyć wektor wartości średnich oraz macierz kowariancji ($S_Y$) dla $Y$.

### Realizacja

Za pomocą funkcji: *cov_matrix*, *standard_deviation* oraz *standardization* dokonano przekształceń. Wyznaczono wektor wartości średnich oraz macierz kowariancji dla **X**. Dane **X** zostały wystandaryzowane (ozn. **Y**). Utworzono macierz kowariancji na podstawie **Y**. Macierz kowariancji dla **Y** jest symetryczna i zawiera na przekątnej wartości 1.

## 1.3 Zadanie I.3 (Spektralna dekompozycja $\mathbf{S_Y}$)

### Polecenie

Wyznaczyć wartości własne i wektory własne macierzy kowariancji $\mathbf{S_Y}$ . Zapisać je jako zmienne *eval* i *evec* odpowiednio.

### Realizacja

Za pomocą wbudowanej funkcji podstawowej *eigen* wyznaczono wartości i wektory własne dla $\mathbf{Y}$.

## 1.4 Zadanie I.4 (Analiza dekompozycji.)

### Polecenie

Nasze składowe są posortowane względem malejącej wariancji – ponieważ nasze dane były wystandaryzowane, to wariancja $i$-tej składowej równa jest odpowiedniej wartości własnej $\lambda_i$. Przedstawić wypełnioną tabelę z rys.X.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | ... |
|---|---|---|---|---|---|---|---|---|---|---|
| wariancja | · | · | · | · | · | · | · | · | · | · |
| proporcja wariancji ($\tau_i$) | · | · | · | · | · | · | · | · | · | · |
| skumulowana proporcja | · | · | · | · | · | · | · | · | · | · |

Rysunek 1: Docelowa tabela

### Realizacja

W poniższej tabeli umieszczono wyniki wariancji (utożsamianej z wartościami własnymi), proporcję wariancji i skumulowaną proporcję. Można zauważyć, że pierwsza składowa główna wyjaśnia 44,27% wariancji, a trzy pierwsze składowe główne 72,64%.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| wariancja | 1.328161e+01 | 5.691355e+00 | 2.817949e+00 | 1.980640e+00 | 1.648731e+00 | 1.207357e+00 |
| proporcja wariancji | 4.427203e-01 | 1.897118e-01 | 9.393163e-02 | 6.602135e-02 | 5.495768e-02 | 4.024522e-02 |
| skumulowana proporcja | 0.4427203 | 0.6324321 | 0.7263637 | 0.7923851 | 0.8473427 | 0.8875880 |
| | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 |
| wariancja | 6.752201e-01 | 4.766171e-01 | 4.168948e-01 | 3.506935e-01 | 2.939157e-01 | 2.611614e-01 |
| proporcja wariancji | 2.250734e-02 | 1.588724e-02 | 1.389649e-02 | 1.168978e-02 | 9.797190e-03 | 8.705379e-03 |
| skumulowana proporcja | 0.9100953 | 0.9259825 | 0.9398790 | 0.9515688 | 0.9613660 | 0.9700714 |
| | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 |
| wariancja | 2.413575e-01 | 1.570097e-01 | 9.413497e-02 | 7.986280e-02 | 5.939904e-02 | 5.261878e-02 |
| proporcja wariancji | 8.045250e-03 | 5.233657e-03 | 3.137832e-03 | 2.662093e-03 | 1.979968e-03 | 1.753959e-03 |
| skumulowana proporcja | 0.9781166 | 0.9833503 | 0.9864881 | 0.9891502 | 0.9911302 | 0.9928841 |
| | PC19 | PC20 | PC21 | PC22 | PC23 | PC24 |
| wariancja | 4.947759e-02 | 3.115940e-02 | 2.997289e-02 | 2.743940e-02 | 2.434084e-02 | 1.805501e-02 |
| proporcja wariancji | 1.649253e-03 | 1.038647e-03 | 9.990965e-04 | 9.146468e-04 | 8.113613e-04 | 6.018336e-04 |
| skumulowana proporcja | 0.9945334 | 0.9955720 | 0.9965711 | 0.9974858 | 0.9982971 | 0.9988990 |
| | PC25 | PC26 | PC27 | PC28 | PC29 | PC30 |
| wariancja | 1.548127e-02 | 8.177640e-03 | 6.900464e-03 | 1.589338e-03 | 7.488031e-04 | 1.330448e-04 |
| proporcja wariancji | 5.160424e-04 | 2.725880e-04 | 2.300155e-04 | 5.297793e-05 | 2.496010e-05 | 4.434827e-06 |
| skumulowana proporcja | 0.9994150 | 0.9996876 | 0.9999176 | 0.9999706 | 0.9999956 | 1.0000000 |

## 1.5 Zadanie I.5 (Interpretacja składowych głównych.)

### Polecenie

Zauważmy, że najbardziej informatywną kombinacją liniową zmiennych jest ta zadana przez pierwszy wektor własny (odpowiadający największej wartości własnej). Zapisz w postaci liniowej kombinacji wycentrowanych zmiennych $y_1, ..., y_{30}$ pierwszą i drugą składową główną $z_1$, $z_2$. Czy potrafisz opisać / wytłumaczyć w praktyce co opisuje zmienna $z_1$, a co $z_2$?

Przedstawić na wykresie zdolność kolejnych składowych głównych do wyjaśniania zmienności w danych. W tym celu wykreślić wykres osypiska (ang. *scree plot*) opisujący wyjaśnianą wariancję – względnie lub bezwzględnie.

### Realizacja

Dla $z_1$ przyjmujemy za współczynniki elementy wektora własnego dla odpowiednich zmiennych

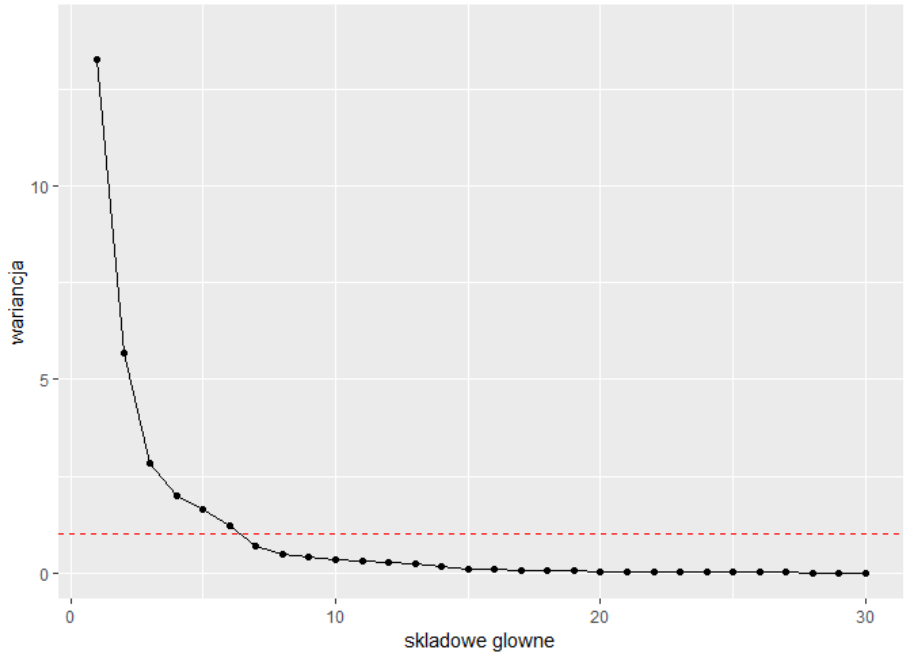
$$\begin{aligned} z_1 =& 0.2189 \cdot y_1 - 0.1037 \cdot y_2 - 0.2275 \cdot y_3 - 0.2210 \cdot y_4 - 0.1426 \cdot y_5 + \\ & - 0.2393 \cdot y_6 - 0.2584 \cdot y_7 - 0.2609 \cdot y_8 - 0.1382 \cdot y_9 - 0.0644 \cdot y_{10} + \\ & - 0.2060 \cdot y_{11} - 0.01742 \cdot y_{12} - 0.2113 \cdot y_{13} - 0.2029 \cdot y_{14} - 0.2029 \cdot y_{15} + \\ & - 0.1704 \cdot y_{16} - 0.1536 \cdot y_{17} - 0.1834 \cdot y_{18} - 0.04250 \cdot y_{19} - 0.1026 \cdot y_{20} + \\ & - 0.2280 \cdot y_{21} - 0.1044 \cdot y_{22} - 0.2366 \cdot y_{23} - 0.2259 \cdot y_{24} - 0.1280 \cdot y_{25} + \\ & - 0.2101 \cdot y_{26} - 0.2288 \cdot y_{27} - 0.2509 \cdot y_{28} - 0.1229 \cdot y_{29} - 0.1318 \cdot y_{30}. \end{aligned}$$

Można stwierdzić, że im mniejsza wartość danego współczynnika, tym mniejsza jest istotność cechy, której odpowiada.

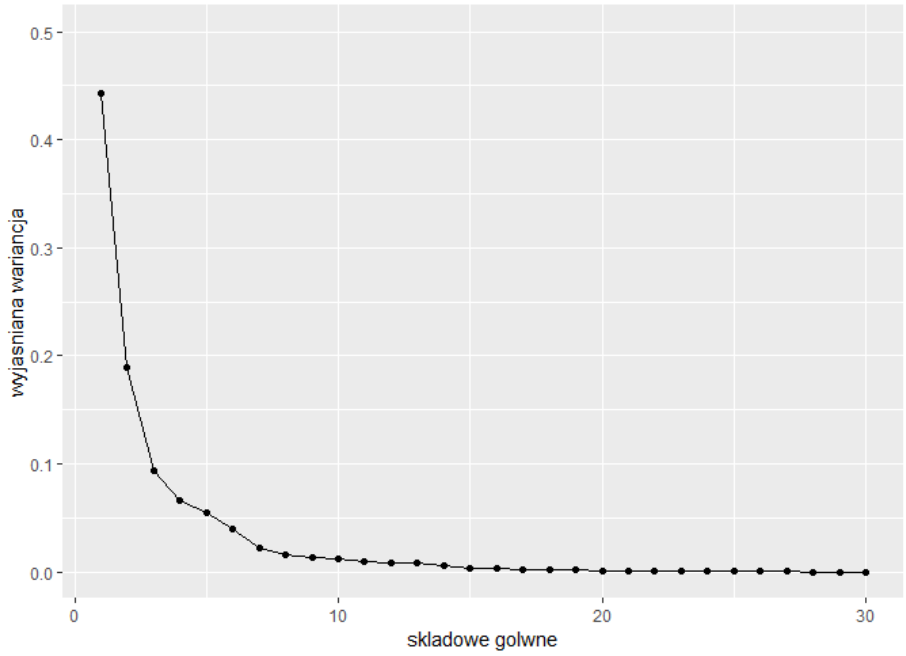
Formuła na składową drugą:

$$\begin{aligned} z_2 =& - 0.2339 \cdot y_1 - 0.0597 \cdot y_2 - 0.2152 \cdot y_3 - 0.2311 \cdot y_4 + 0.1861 \cdot y_5 + \\ & - 0.1519 \cdot y_6 - 0.0602 \cdot y_7 - 0.0358 \cdot y_8 + 0.1903 \cdot y_9 + 0.3666 \cdot y_{10} + \\ & - 0.1056 \cdot y_{11} + 0.0900 \cdot y_{12} - 0.0895 \cdot y_{13} - 0.1523 \cdot y_{14} + 0.2044 \cdot y_{15} + \\ & + 0.2327 \cdot y_{16} + 0.1972 \cdot y_{17} + 0.1303 \cdot y_{18} + 0.1838 \cdot y_{19} + 0.2801 \cdot y_{20} + \\ & - 0.2199 \cdot y_{21} - 0.0455 \cdot y_{22} - 0.1999 \cdot y_{23} - 0.2194 \cdot y_{24} + 0.1723 \cdot y_{25} + \\ & + 0.1436 \cdot y_{26} + 0.0980 \cdot y_{27} - 0.0083 \cdot y_{28} + 0.1419 \cdot y_{29} + 0.2753 \cdot y_{30}. \end{aligned}$$

Na rys. 2 i rys. 3 przedstawiono wykresy osypiska opisujące wariancję względnie i bezwzględnie.




Rysunek 2: Wykres osypiska opisujący wariancję.




Rysunek 3: Wykres osypiska opisujący wyjaśnianą wariancję.

Składowe o wartości wariancji mniejszej od 1 mają mniejszą wariancję niż oryginalne dane.

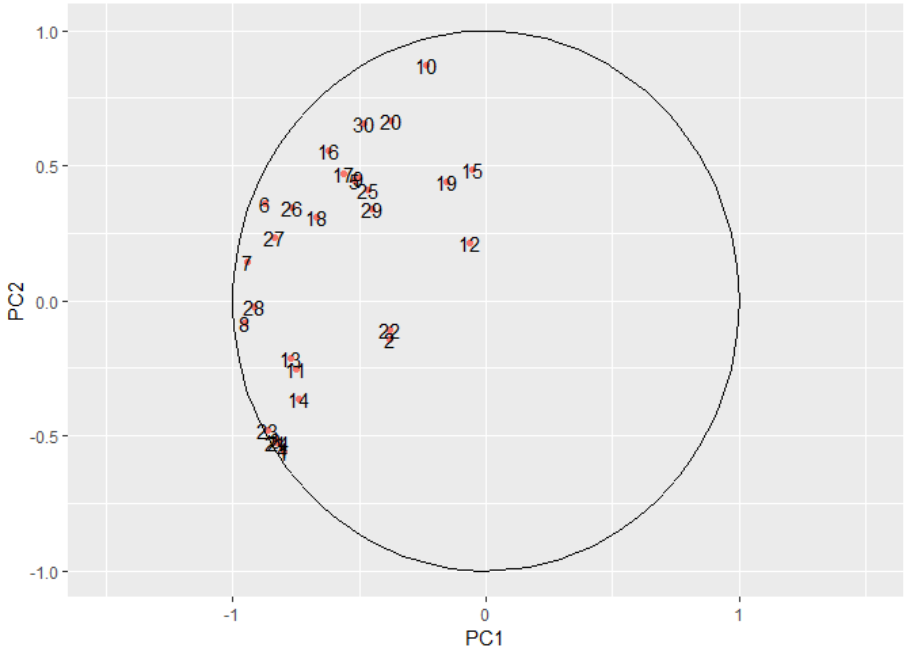
## 1.6 Zadanie I.6 (Korelacja pomiędzy zmiennymi a składowymi głównymi.)

### Polecenie

Rozważmy korelację pomiędzy wektorem składowych głównych $\mathbf{Z}$ i oryginalnym, wystandaryzowanym wektorem $\mathbf{Y}$ . Korelacje $r_{Y_i,Z_j}$ mogą być wykorzystane do określenia związku pomiędzy składową główną $Z_j$ i oryginalną zmienną $X_i$. Zauważmy, że $\sum_{j=1}^{p} r^2_{Y_i,Z_j} = 1$, dla każdego $i$. Stąd $r^2_{Y_i,Z_j}$ mogą być traktowane jako proporcja wariancji $Y_i$ wyjaśniana przez $Z_j$ . Wykreślić powyższe proporcje w przestrzeni pierwszych dwóch składowych głównych.

### Realizacja

Na rys. 4 przedstawiono wykres proporcji w przestrzeni pierwszych dwóch składowych głównych.




Rysunek 4: Wykres proporcji w przestrzeni PC1 i PC2.

Im zmienne leżą bliżej okręgu, tym są lepiej wyjaśniane przez składowe PC1 i PC2, oznacza to również, że są z nimi najsilniej skorelowane. Z rys. 4 można odczytać te zmienne, są to np. $Y_{24}, Y_{23}$.

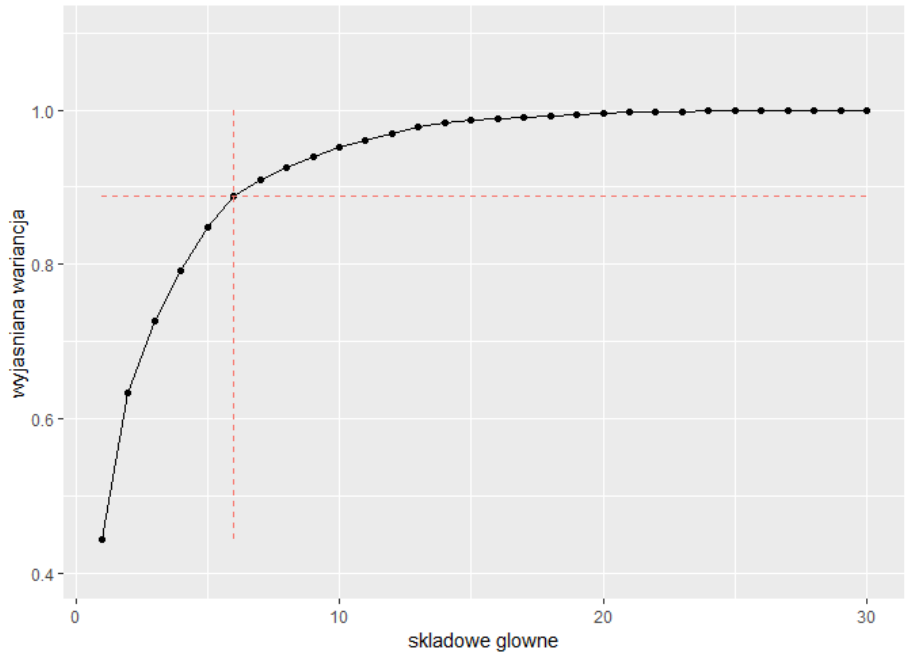
## 1.7 Zadanie I.7 (Wybór składowych głównych w celu redukcji wymiaru.)

### Polecenie

Wybrać na podstawie wykresu osypiska składowe główne, dla których wariancja jest większa od 1. Wykreślić wykres skumulowanej proporcji i określić wartość wyjaśnianej wariancji przez te składowe główne.

### Realizacja

Na rys. 5 przedstawiono wykres skumulowanej proporcji.




Rysunek 5: Wykres skumulowanej proporcji.

Znormalizowane składowe główne, dla których wariancja jest mniejsza niż 1 w praktyce wyjaśniają mniejszą część wariancji niż oryginalne zmienne (ich wariancja wynosiła 1). Zauważamy, że po składowej 6 wykres zaczyna się spłaszczać, czyli maleje wzrost wyjaśnialności składowych. Zatem w wyniki kompromisu pomiędzy wyjaśnialnością, a liczbą zmiennych nie powinno się uwzględniać więcej czynników, niż te znajdujące się do 6 punktu. Zatem moglibyśmy użyć tylko sześciu zmiennych do opisu danych. Zysk naszej analizy to redukcja aż 24 wymiarów. W kontekście wyłącznie analizy PCA surowe dane $\mathbf{X}$ nie są już potrzebne, jednakże PCA służy lepszemu poglądowi na dane, zatem w szerszym kontekście dalszych analiz (np. tworzenia modeli itp.) surowe dane są jak najbardziej potrzebne.

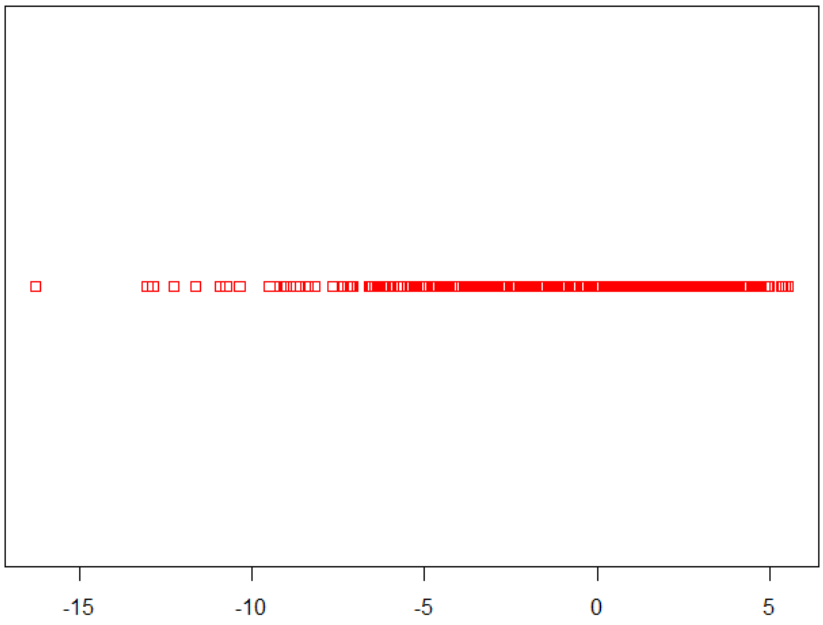
## 1.8 Zadanie I.8 (Wykresy składowych głównych.)

### Polecenie

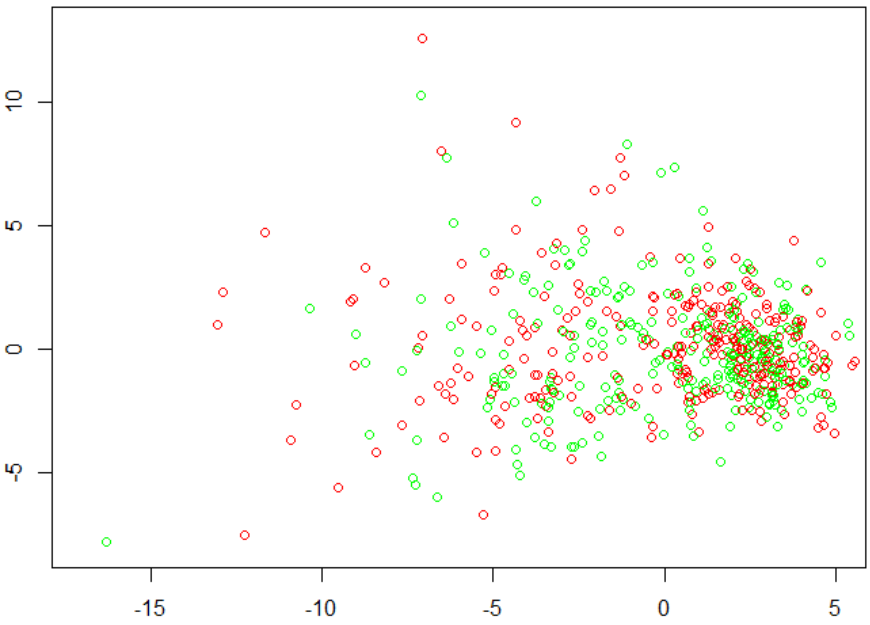
Przedstaw jednowymiarową, dwuwymiarową i trójwymiarową projekcje na PC1, PC2, PC3.

### Realizacja

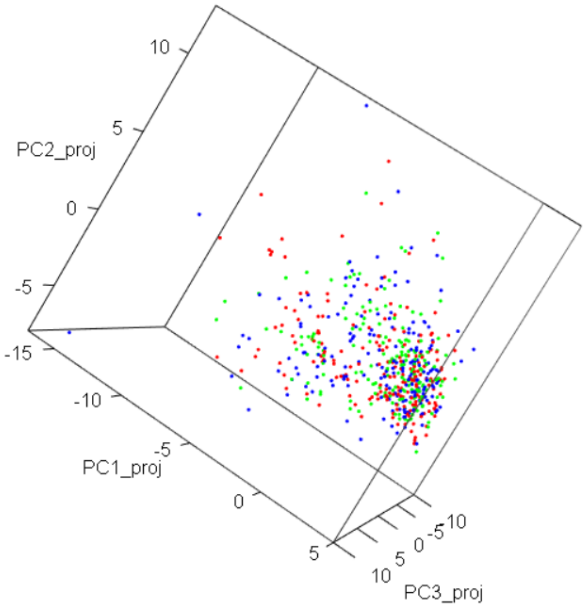
Na poniższych rysunkach przedstawiono kolejno: jednowymiarową projekcję, dwuwymiarową projekcję i trójwymiarową projekcję.




Rysunek 6: Jednowymiarowa projekcja.




Rysunek 7: Dwuwymiarowa projekcja.




Rysunek 8: Trójwymiarowa projekcja.

Można zaobserwować, że dane są skupione w jednej części wykresu. Możemy również zaobserwować, że istnieją obserwacje odstające.

## 1.9 Zadanie I.9 (Dyskryminacja.)

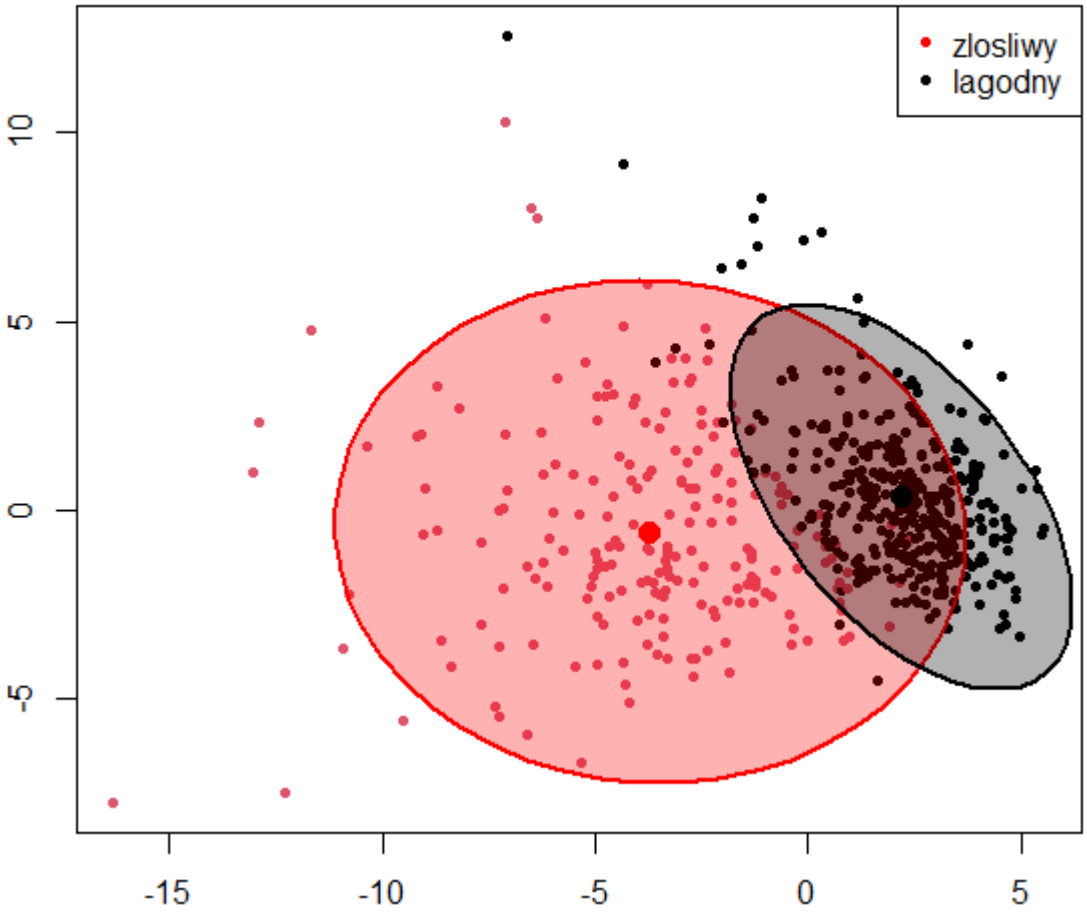
### Polecenie

Załóżmy, że nasza analiza ma posłużyć wskazaniu / wytłumaczeniu / opisaniu różnic pomiędzy złośliwym a łagodnym guzem. Dodaj zmienną odpowiedzi (diagnosis) do wykresów i zinterpretuj je.

Przedstaw projekcję na PC1 i PC2 obserwacji w podziale na złośliwy / łagodny. Wyznacz środki obu zbiorów i zaznacz (wyznacz elipsy) zbiory zawierające najbardziej typowe 95% obserwacji z każdego z podzbiorów.

### Realizacja

Dokonano podziału na obserwacje złośliwe (czerwone) i łagodne (czarne) oraz wyznaczono elipsy określające te dwa podziały. Wyniki umieszczono na rys. 9.




Rysunek 9: Podział na nowotwory łagodne i złośliwe.

Możemy zauważyć, że obserwacje układają się w dwóch obszarach. Obserwacje złośliwe są bardziej rozproszone (stąd większe pole elipsy) niż łagodne, które charakteryzują się większym zagęszczeniem obserwacji. Można również postawić zasadną hipotezę, że punkty znajdujące się w dużej odległości od elipsy to obserwacje odstające (lub w pewien sposób błędne).

# 2 Część druga

## Opis problemu

W tej części posłużymy się gotowym pakietem statystycznym wbudowanym w R, aby powtórzyć tą samą analizę co w części pierwszej.

Nie będziemy analizować wyników, a jedynie porównamy je z tymi uzyskanymi w części pierwszej.

## 2.1 Zadanie II.1 (Wczytanie danych.)

### Realizacja

Identycznie jak w 1.1, zmieniono jedynie nazwy zmiennych.

## 2.2 Zadanie II.2 (Przygotowanie danych i statystyki opisowe.)

### Realizacja

Skorzystano z wbudowanych funkcji *colMeans, cov* oraz *prcomp*. Uzyskane wyniki w przybliżeniu pokrywają się z uzyskanymi 1.2. Funkcja *prcomp* przyjmuje jako argumenty dane, a także argumenty boolowskie pozwalające na standaryzację i centrowanie zmiennych.

## 2.3 Zadanie II.3 (Spektralna dekompozycja $\mathbf{S_Y}$)

### Realizacja

Identycznie jak w 1.3.

## 2.4 Zadanie II.4 (Analiza dekompozycji.)

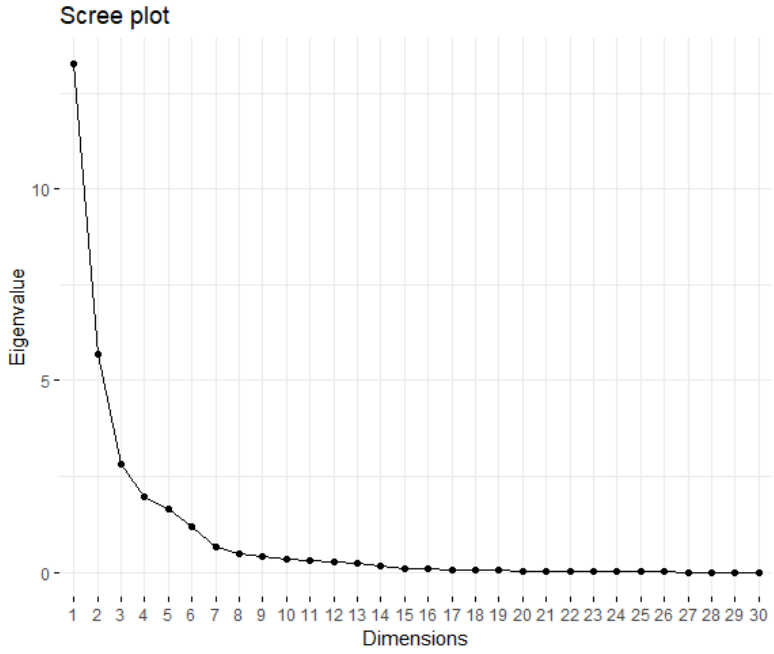
### Realizacja

Skorzystano z funkcji *get_eigenvalue*. Uzyskane wyniki w przybliżeniu są takie same jak te uzyskane w 1.4. Funkcja *get_eigenvalue* przyjmuje za argument wynik działania funkcji *prcomp* i zwraca wartość własną (wariancję), procent proporcji wariancji oraz procent skumulowanej proporcji.

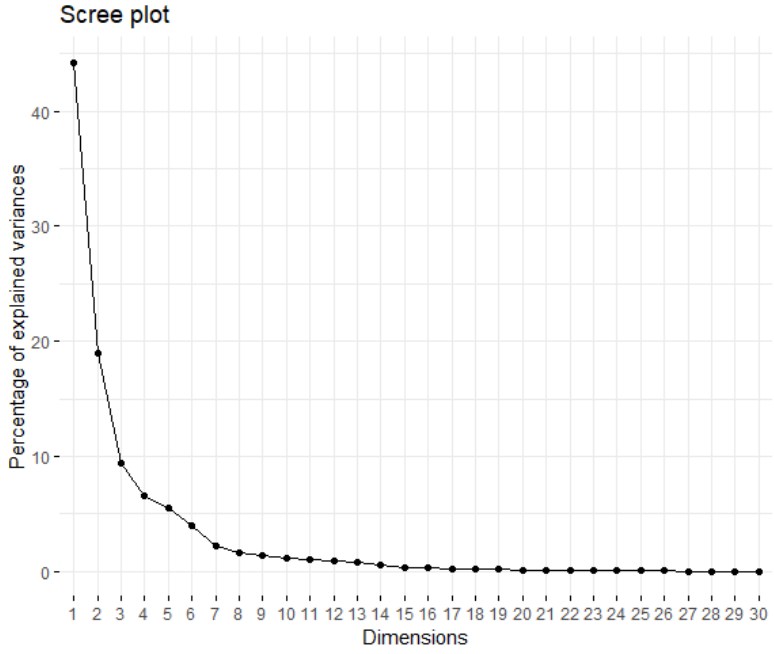
## 2.5 Zadanie II.5 (Interpretacja składowych głównych.)

### Realizacja

Na podstawie wyników funkcji z 2.4 widać, że wyniki są identyczne do tych z 1.5. Na poniższych rysunkach umieszczono wykresy osuwiska utworzone za pomocą funkcji *fviz_eig*.




Rysunek 10: Wykres osypiska opisujący wariancję.



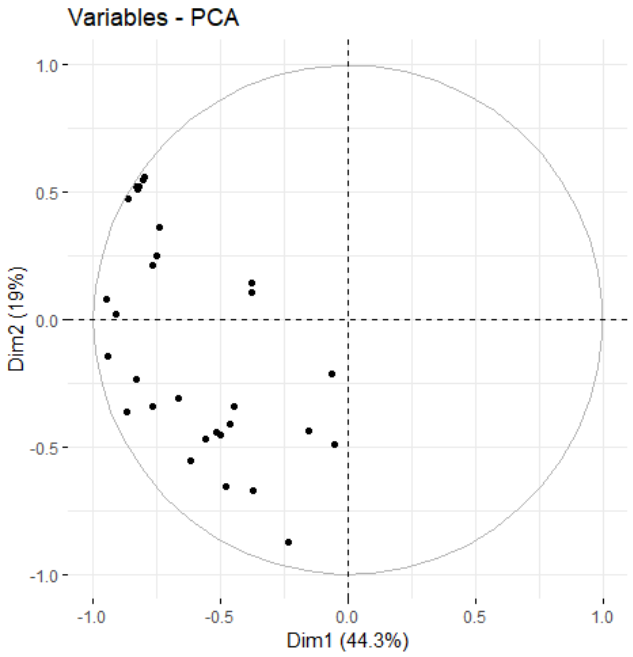

Rysunek 11: Wykres osypiska opisujący procent wyjaśnianej wariancji.

Porównując z wykresami z 1.5 można stwierdzić, że wyniki są bardzo zbliżone. Funkcja *fviz_eig* przyjmuje za argument wynik PCA oraz typ wykresu. Można uzyskać wykres osypiska opisujący wariancję albo wykres osypiska opisujący procent wyjaśnianej wariancji.

## 2.6 Zadanie II.6 (Korelacja pomiędzy zmiennymi a składowymi głównymi.)

### Realizacja

Posłużono się funkcją *fviz_pca_var*, której wynik umieszczono na poniższym rysunku. Jak widać, porównując z 1.6 punkty są odbite względem osi Dim2, poza tym



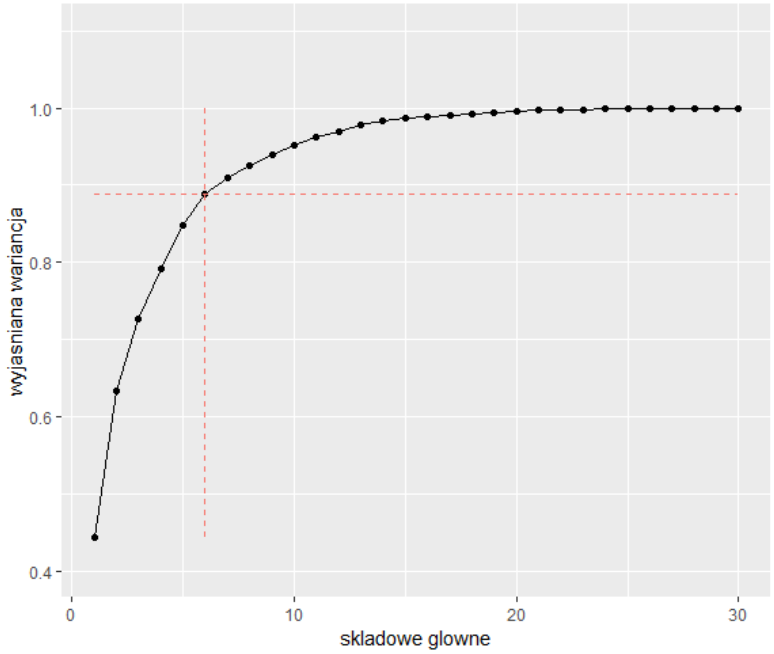

Rysunek 12: Wykres proporcji w przestrzeni PC1 i PC2.

umiejscowienie jest identyczne. Dodatkowo na osiach widać procent wyjaśnialności. Funkcja *fviz_pca_var* tworzy wykres proporcji w przestrzeni PC1 i PC2.

## 2.7 Zadanie II.7 (Wybór składowych głównych w celu redukcji wymiaru.)

### Realizacja

Skorzystano z tego samego kodu co w 1.7, zmieniając jedynie zmienną na tę z 2.2.



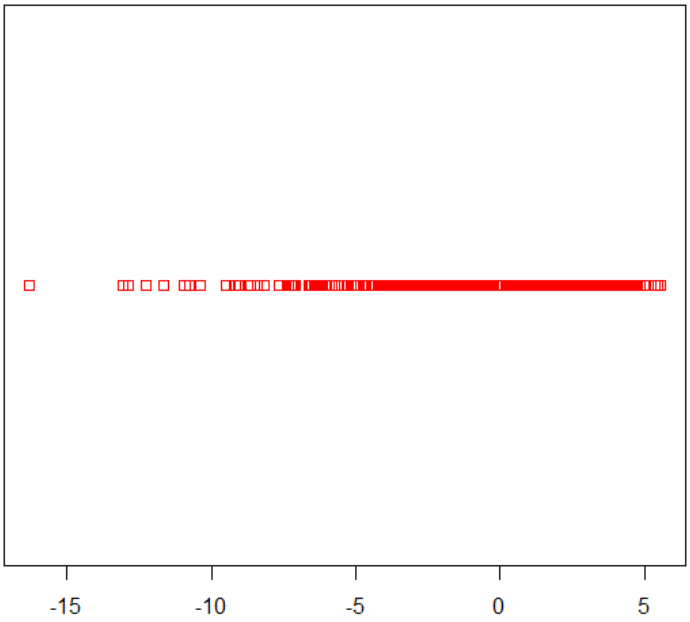

Rysunek 13: Wykres skumulowanej proporcji.

Wykres jest identyczny jak w 1.7.

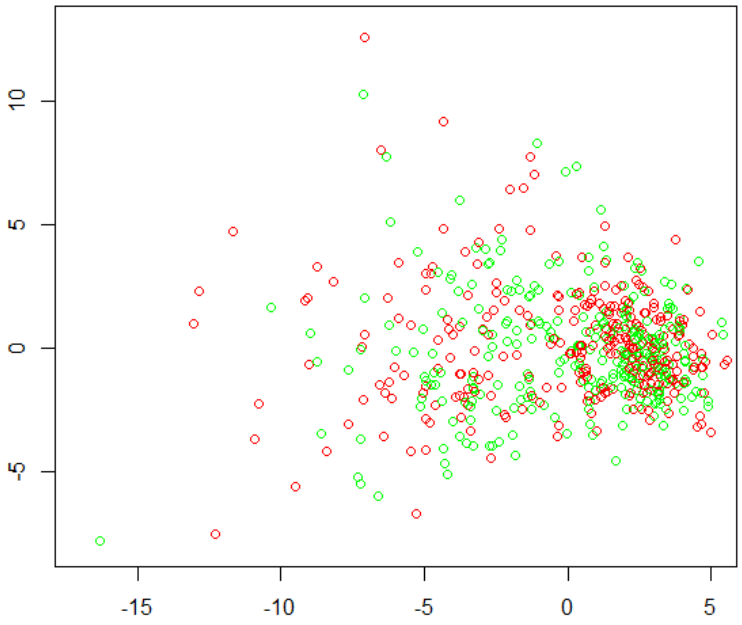
## 2.8 Zadanie II.8 (Wykresy składowych głównych.)

### Realizacja

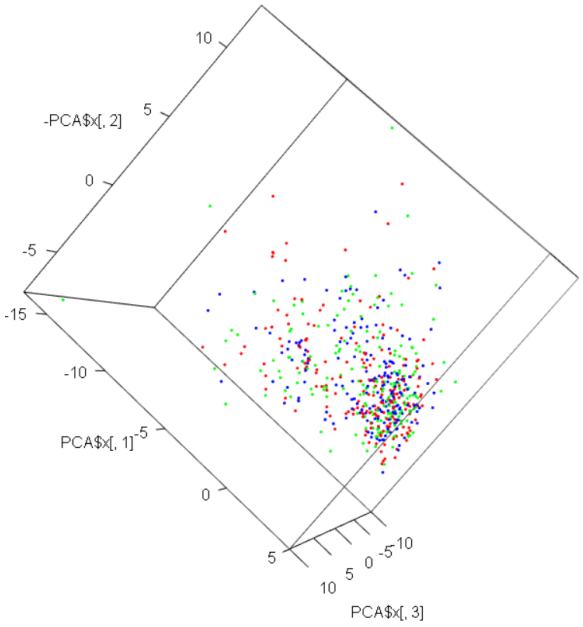
Identyczny kod jak w 1.8 z danymi z 2.2. Wartość PC2 wzięto z minusem w celu uzyskania identycznych wyników.




Rysunek 14: Jednowymiarowa projekcja.




Rysunek 15: Dwuwymiarowa projekcja.



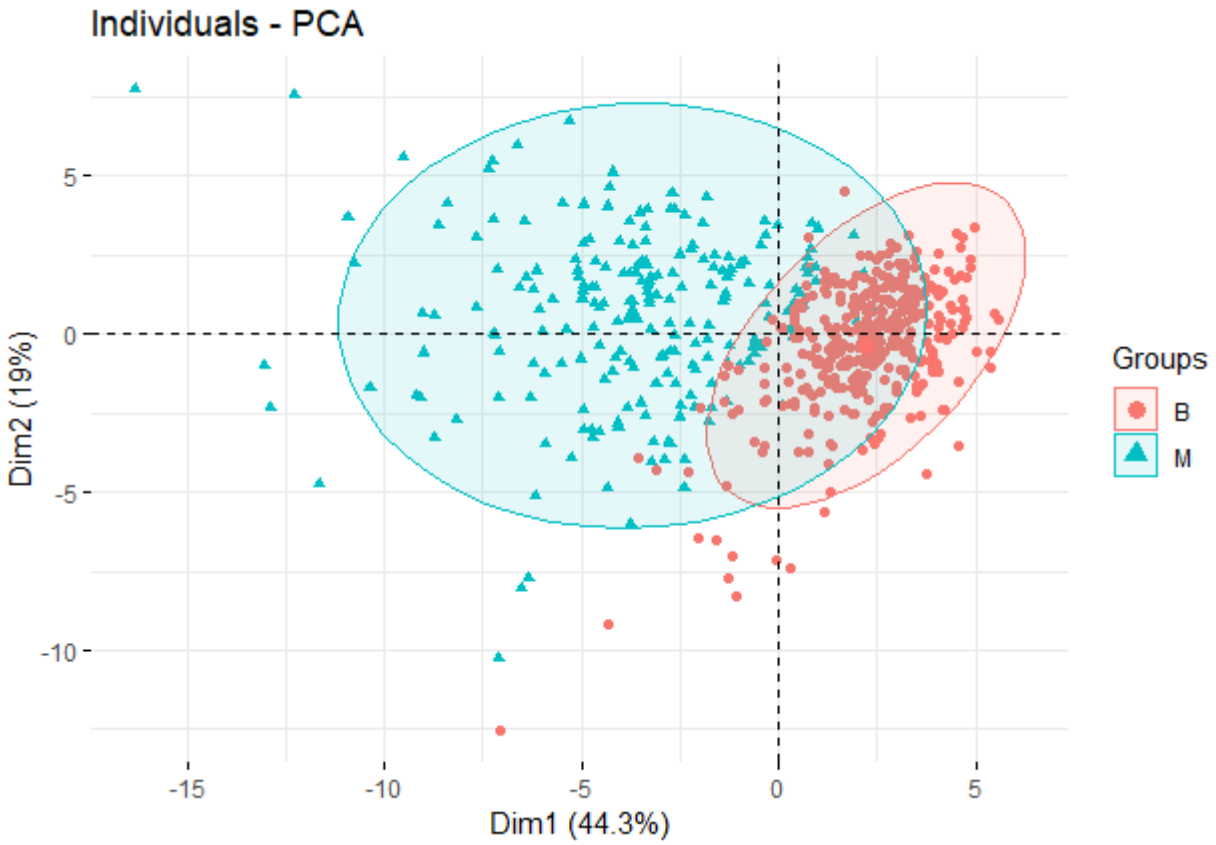

Rysunek 16: Trójwymiarowa projekcja.

Wykresy są identyczne jak 1.8.

## 2.9 Zadanie II.9 (Dyskryminacja.)

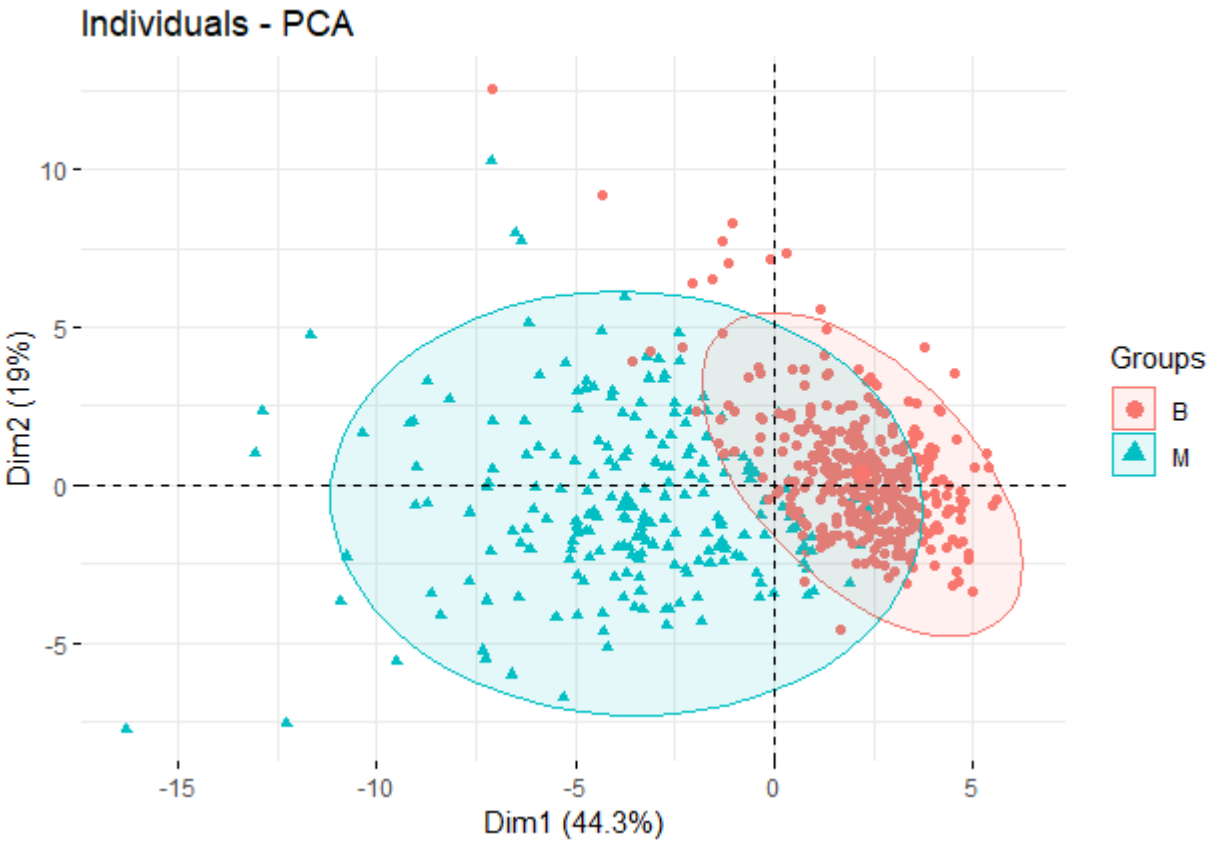
### Realizacja

Korzystając z funkcji *fviz_pca_ind* wykreślono elipsy z rys. 17




Rysunek 17: Podział na nowotwory łagodne i złośliwe.

Rozbieżność z 1.9 wynika z przeciwnych znaków dla PC2.




Rysunek 18: Skorygowany podział na nowotwory łagodne i złośliwe.

## 2.10 Wniosek

Porównując wyniki wyprowadzone w cz. II można wnioskować, że poprawnie dokonano analizy w cz. I. Mnożąc PC2 przez -1 otrzymano wyniki identyczne.