METODA REGRESJI GŁÓWNYCH SKŁADOWYCH (PCR)

Istotnym ograniczeniem stosowania MLR w modelowaniu QSAR jest to, że gdy pomiędzy zmiennymi występują silne korelacje nie jest możliwe poprawne odwrócenie macierzy ($\boldsymbol{X}^T\boldsymbol{X}$), a więc wzór nie może zostać użyty do obliczenia współczynników $\boldsymbol{b}$. W tego typu przypadkach konieczne jest skorzystanie z innej metody np. regresji głównych składowych (PCR, *Principal Component Regression*) – zamiast oryginalnych zmiennych objaśniających wykorzystywane są wówczas niezależne od siebie (ortogonalne) główne składowe.

Algorytm PCR składa się z trzech etapów:

1) zastosowanie analizy głównych składowych(PCA) do wygenerowania głównych składowych,
2) zachowanie k pierwszych głównych składowych, które wyjaśniają największą ilość wariancji w danych,
3) dopasowanie modelu regresji liniowej (metoda najmniejszych kwadratów) do k głównych składowych.

Analiza głównych składowych

Pierwszym etapem analizy głównych składowych jest utworzenie macierzy korelacji-kowariancji $\boldsymbol{C}$ $(m \times m)$ na podstawie autoskalowanej macierzy danych $\boldsymbol{X}$ $(n \times m)$ zgodnie ze wzorem (4):

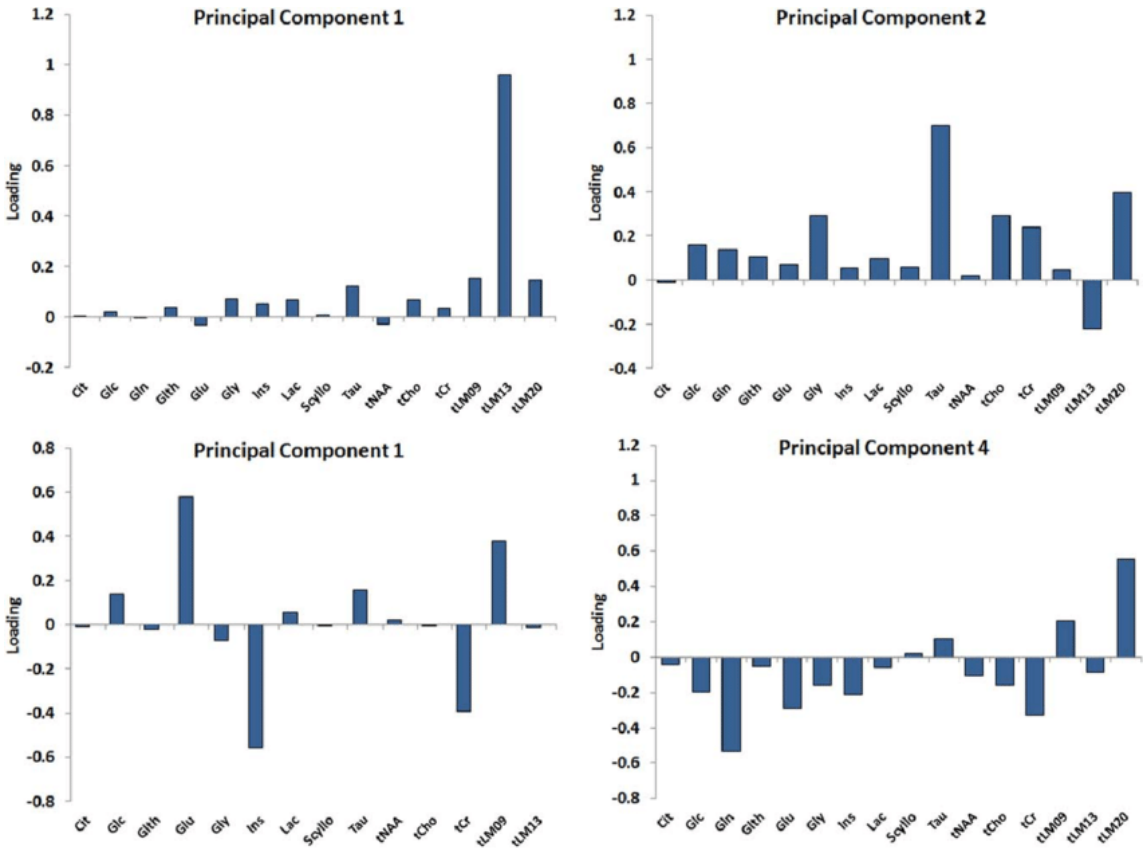
$$\boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{X} \tag{1}$$

Następnie wyznacza się wektory własne macierzy **C** (macierz **W**). Elementy wektorów własnych są współczynnikami kombinacji liniowej zmiennych objaśniających definiujących poszczególne główne składowe. Z każdym z wektorów własnych związana jest jedna wartość własna λ_i. Liczba ta charakteryzuje zasób informacji (zmienności) wyjaśnianej przez daną zmienną.

PCA zakłada, że zmienność właściwa uwzględniania jest w k pierwszych głównych składowych o największych wartościach własnych, przy czym wartości własne są proporcjonalne do ilości wyjaśnianej informacji.

W następnym kroku dla wybranych głównych składowych obliczane są dwie macierze: macierz ładunków czynnikowych $\boldsymbol{P}$ oraz macierz wartości czynnikowych $\boldsymbol{T}$.

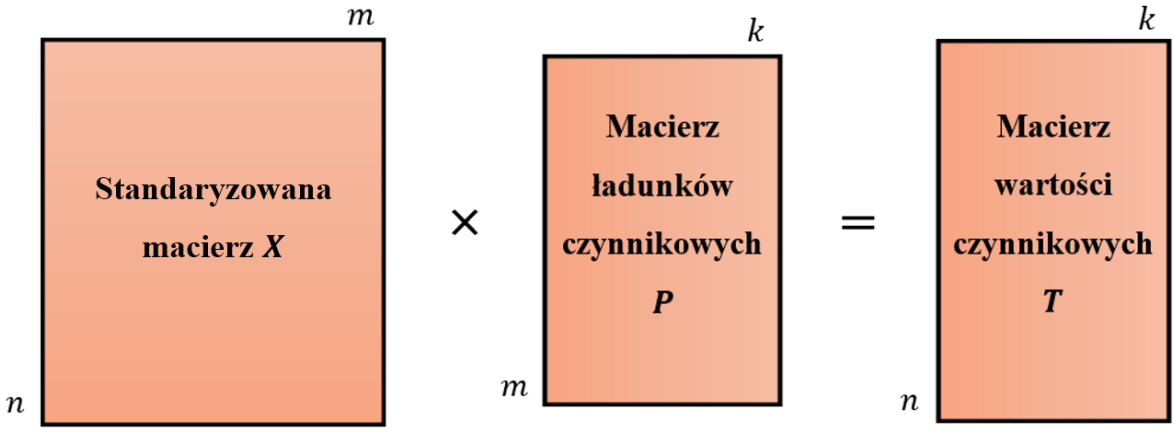
Macierz $\boldsymbol{P}$ o wymiarach $n \times k$ otrzymuje się poprzez odcięcie z macierzy $\boldsymbol{C}$ wektorów nieistotnych głównych składowych. Jej elementy stanowią ładunki wnoszone do kolejnych składowych przez poszczególne zmienne. Innymi słowy, macierz ta opisuje zależności między zmiennymi w przestrzeni głównych składowych. Zgodnie z regułą Malinowskiego istotne są te zmienne, których znormalizowane wartości ładunków czynnikowych są większe lub równe 0,7, lub mniejsze bądź równe -0,7. Informację tę można przedstawić graficznie w formie wykresu ładunków czynnikowych.



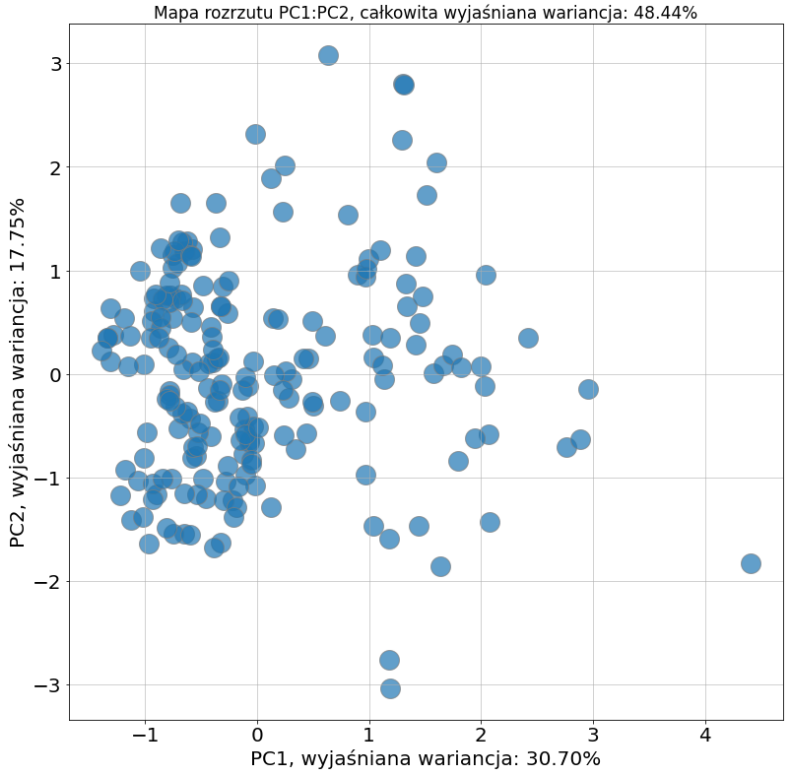
Rysunek 1. Przykładowy wykres ładunków czynnikowych dla 3 głównych składowych

Na podstawie wykresu ładunków czynnikowych można ocenić jaki wkład i jak skorelowane są poszczególne zmienne objaśniające w tworzenie głównych składowych.

Macierz $\boldsymbol{T}$ o wymiarach $m \times k$ powstaje w wyniku pomnożenia autoskalowanej macierzy $\boldsymbol{X}$ przez macierz $\boldsymbol{P}$ (**Rysunek 2.**) i zawiera współrzędne obiektów w przestrzeni nowych składowych (zmiennych). Na jej podstawie tworzy się tzw. mapy liniowe przedstawiające rzuty przestrzeni na płaszczyznę wyznaczaną przez kolejne główne składowe. Przykładowa mapa liniowa przedstawiona jest na **Rynsuku 3**.



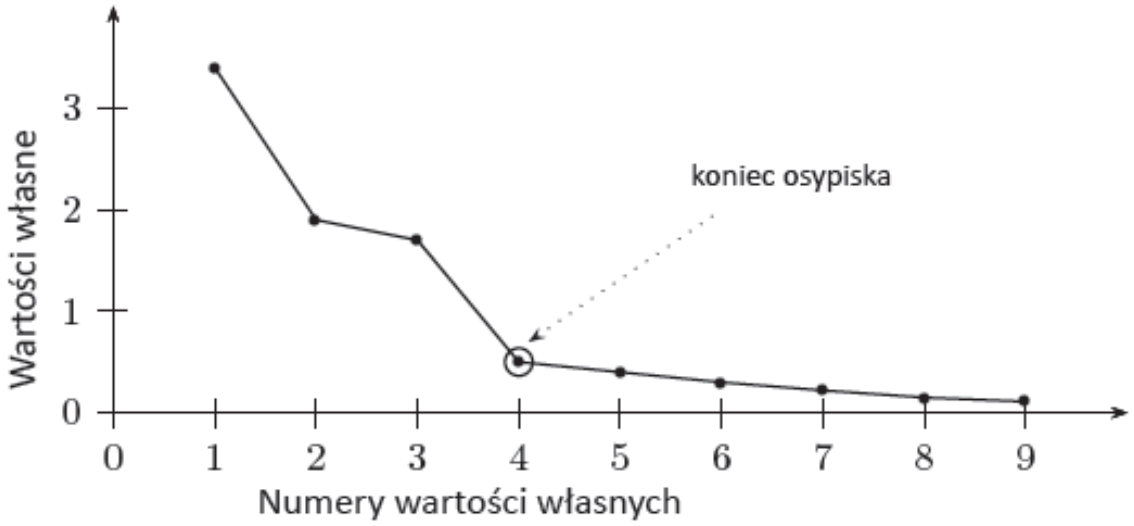
Rysunek 2. Schemat przekształceń prowadzący do uzyskania współrzędnych obiektów w wielowymiarowej przestrzeni cech.



Rysunek 3. Przykładowa mapa liniowa PC1:PC2

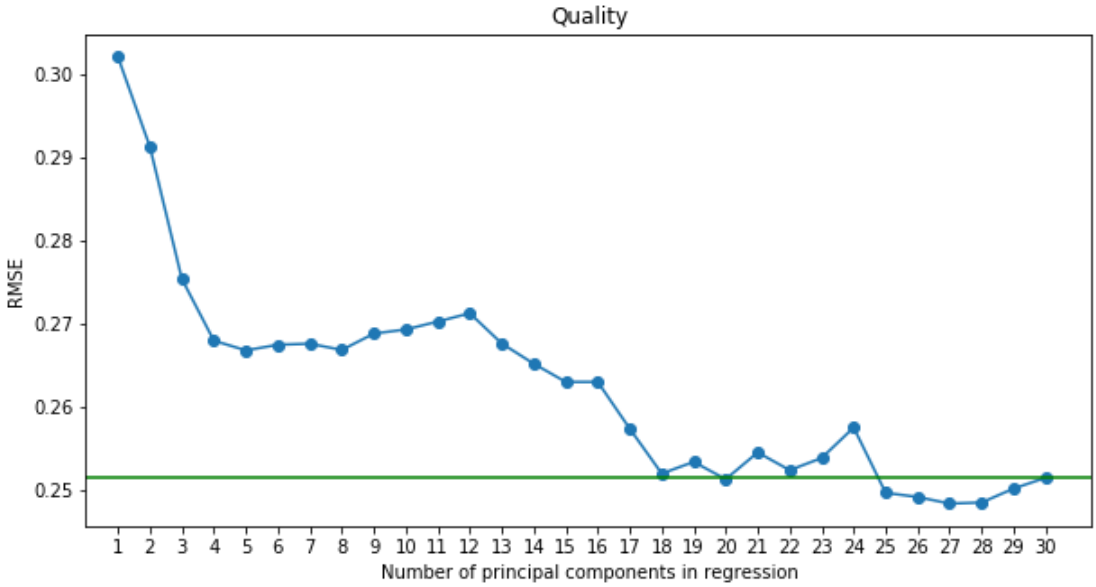
Optymalną liczbę głównych składowych możemy wyznaczyć na podstawie kilku kryteriów:

1. Kryterium Kaisera – wykorzystywane są główne składowe których wartości własne są większe niż 1.
2. Wykres osypiska – liczbę głównych składowych wyznacza się na podstawie punktu w którym wykres zaczyna się wypłaszczać i nie ma gwałtownych spadków wartości własnej. Oznacza to, że po tym punkcie nie ma znaczącego przyrostu informacji.



Rysunek 4. Przykładowy wykres osypiska

3. Sumaryczny procent wyjaśnianej wariancji – umowne określenie, ile procent maksymalnej wariancji wyjaśnianej przez główne składowe jest wystarczające. Najczęściej przyjmuje się, że akceptowalny poziom wynosi 70-80%.
4. Średni błąd kwadratowy – metoda wykorzystywana, jeśli na otrzymanych składowych planowana jest dalsza regresja. Buduje się model regresji i wyznacza wartość RMSE z każdą dodaną główną składową. Wybiera się taką ilość głównych składowych dla których wynik RMSE jest najniższy.



Rysunek 5. Przykładowy wykres RMSE od ilości głównych składowych.