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 do wygenerowania głównych składowych,
2) zachowanie k pierwszych głównych składowych, które wyjaśniają największą ilość wariancji w danych (k jest determinowane przez walidację krzyżową),
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):

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

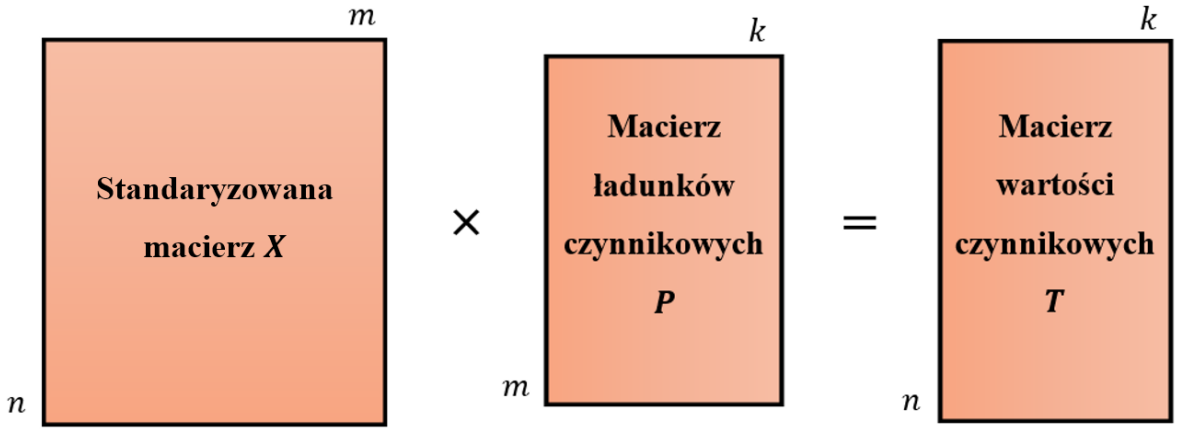
Następnie wyznacza się wektory własne macierzy **C** (macierz **W**). Elementy wektorów własnych są współczynnikami kombinacji liniowej zmiennych objaśniających definiujących poszczególne główne składowe (czynniki). Z każdym z wektorów własnych związana jest jedna wartość własna λ_i. Liczba ta charakteryzuje zasób informacji (zmienności) wyjaśnianej przez daną zmienną. Suma wartości własnych odpowiada liczbie zmiennych objaśniających, ponieważ każda z nich ma wariancję równą 1.

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 $m \times k$ otrzymuje się poprzez odcięcie z macierzy $\boldsymbol{C}$ wektorów nieistotnych głównych składowych. Jej elementy stanowią ładunki wnoszone do kolejnych czynników 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.

Macierz $\boldsymbol{T}$ o wymiarach $n \times k$ powstaje w wyniku pomnożenia autoskalowanej macierzy $\boldsymbol{X}$ przez macierz $\boldsymbol{P}$ (**Rysunek 1.**) i zawiera współrzędne obiektów w przestrzeni nowych czynników (zmiennych). Na jej podstawie tworzy się tzw. mapy liniowe przedstawiające rzuty przestrzeni na płaszczyznę wyznaczaną przez kolejne główne składowe.



Rysunek 1. Schemat przekształceń prowadzący do uzyskania współrzędnych obiektów w wielowymiarowej przestrzeni cech.

ZADANIE. Proszę zbudować model metodą PCR wg poniższej instrukcji:

1) Import standaryzowanego zestawu danych dane_leki.xlsx
2) Przeprowadzenie analizy PCA (biblioteka: sklearn.decomposition.pca)
3) Podział na zbiór uczący i walidacyjny (biblioteka: sklearn.model_selection):
   - train_test_split → test_size = 0.33, random_state = 42
4) Wykreślenie zależności $RMSE_C$ od liczby uwzględnianych głównych składowych:
   - metoda walidacji krzyżowej KFold (sklearn.model_selection) → n_splits = 10, shuffle = True, random_state = 0
5) Zbudowanie modelu regresji liniowej dla istotnej liczby głównych składowych
6) Obliczenie R^2 i RMSE oddzielnie dla zbiorów kalibracyjnego i walidacyjnego

Proszę krótko zinterpretować wykres i uzyskane wyniki.