METODA REGRESJI CZĘŚCIOWYCH NAJMNIEJSZYCH KWADRATÓW (PLS)

Regresja częściowych najmniejszych kwadratów (PLS, ang. *Partial Least Squares regression*) to metoda statystyczna stosowana w analizie danych wielowymiarowych, szczególnie w sytuacjach, gdy liczba zmiennych niezależnych (deskryptorów) jest duża w stosunku do obserwacji lub gdy istnieją silne współliniowości między deskryptorami. PLS, podobnie jak w przypadku PCR, wykorzystuje tzw. zmienne ukryte określane również latealnymi (LVs, ang. *latent vectors*). Są one wyodrębniane w oparciu o tzw. macierz iloczynów mieszanych $\boldsymbol{C}$**:**

$$\boldsymbol{C} = \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}$$

W ten sposób uzyskiwany jest efekt maksymalizacji kowariancji pomiędzy macierzą $\boldsymbol{X}$ i wektorem $\boldsymbol{y}$. Pierwszy czynnik opisuje największą część zmienności możliwą do opisania przy użyciu jednej zmiennej ukrytej, drugi największą część pozostałej kowariancji i tak dalej. Poszczególnym czynnikom, podobnie jak w PCR, przypisuje się interpretację fizyczną w oparciu o macierz ładunków czynnikowych.

Kluczowe cechy regresji PLS: (i) radzenie sobie z wysokowymiarową macierzą danych (PLS działa dobrze, gdy liczba zmiennych niezależnych przekracza liczbę obserwacji, co prowadzi do problemów w klasycznej regresji liniowej), (ii) rozwiązywanie problemu współliniowości (eliminacja silnych korelacji pomiędzy zmiennymi), (iii) koncentruje się na maksymalizacji wariancji zmiennych niezależnych pod kątem ich związku ze zmienną zależną (PCR nie uwzględnia zmiennej zależnej).

Proces PLS można opisać w następujących krokach:

1. Tworzenie zmiennych ukrytych (PLS konstuuje nowe zmienne jako kombinacje liniowe oryginalnych zmiennych $\boldsymbol{X}$, maksymalizując kowariancję między nimi a $\boldsymbol{y}$.
2. Regresja na zmiennych ukrytych (algorytm iteracyjny: liczba komponentów może być dobierana na podstawie kryteriów takich jak minimalizacja błędu predykcji lub walidacja krzyżowa).

ZADANIE. Proszę zbudować model PLS wg poniższej instrukcji:

1) Zaimportowanie autoskalowanego zestawu danych. Arkusze Yt oraz Yv zawierają wartości modelowanej wielkości (energia adsorpcji aminokwasu na powierzchni nanocząstki złota) odpowiednio dla związków ze zbioru uczącego i walidacyjnego; arkusze Xt i Xv zawierają deskryptory obliczone odpowiednio dla związków zbioru uczącego i walidacyjnego.
2) Przeprowadzenie modelowania PLS (jedna ukryta zmienna; 3 deskryptory: energia HOMO, polaryzowalność i topologiczny obszar powierzchni).
3) Wykreślenie ypred od yobs z podziałem na zbiór uczący i walidacyjny (legenda).
4) Obliczenie statystyk: $R^2, RMSE_C, Q^2_{CVloo}, RMSE_{CVloo}, Q^2_{Ex}, RMSE_{Ex}$.
5) Zbudować modele MLR i PCR, powtórzyć dla nich kroki 3 i 4., a następnie porównać wyniki uzyskane wszystkimi trzema metodami.