

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 ($X^T X$), a więc wzór nie może zostać użyty do obliczenia współczynników 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:

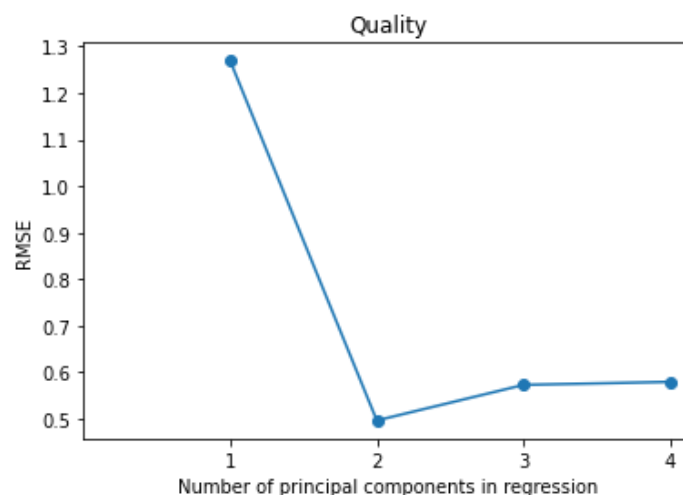
- 1) zastosowanie analizy głównych składowych (PCA) do wygenerowania głównych składowych,
- 2) zachowanie k pierwszych głównych składowych, które wyjaśniają największą ilość wariancji w danych,
- 3) dopasowanie modelu regresji liniowej (metoda najmniejszych kwadratów) do k głównych składowych.

Zmienne uwzględnione w modelu:

1. Zmienna zależna
 - a. parametr $\log K$ HSA – stała równowagowa tworzenia się kompleksu w roztworze; miara siły interakcji między reagentami; wyraża powinowactwo leku do albuminy surowicy człowieka.

Albumina (HSA, ang. human serum albumin) to jedno z dwóch głównych białek osocza odpowiedzialnych za wiązanie leków. Białka osocza są odpowiedzialne za utrzymanie równowagi kwasowo-zasadowej, prawidłowego ciśnienia osmotycznego oraz transport substancji nierozpuszczalnych w wodzie (takich jak endogenne hormony sterydowe lub kwasy tłuszczowe). Czynniki wpływające na stopień wiązania leku z białkami: stężenie leku, powinowactwo leku do białek, oddziaływania cząsteczki leku z „kieszeniami białek”. Lek związany z białkami jest nieaktywny farmakologicznie, nie przenika przez błony biologiczne i nie ulega metabolizmowi, więc zmniejszenie stopnia wiązania leku z białkami osocza skutkuje wzrostem siły działania i skróceniem czasu działania leku.

2. Zmienne niezależne:
 - a. $\log K$ CTAB – retencja (czas retencji = czas uwalniania) w fazie pseudostacjonarnej CTAB (bromek heksadecylotrimetyloamoniowy) z wykorzystaniem metody micelarnej chromatografii elektrokinetycznej (MEKC); micelle utworzone w CTAB mają strukturę podobną do HSA.
 - b. Deskryptory CATS – dostarczają dodatkowych informacji o strukturze cząsteczki oraz mogą dostarczyć użytecznych informacji odzwierciedlających zachowanie leku w regionie wiążącym HSA; kodują informację o częstościach par atomów, które mogą być potencjalnymi miejscami wiązania leku
 - c. CATS3D_09_AL – łączy informacje o lipofilowości i akceptorze wiązań wodorowych.
 - d. CATS3D_00_AA i CATS 3D_00_DD – ważone tylko przez dawcę wiązania wodorowego (D), siłę akceptora (A). Wpływ wiązania wodorowego jest dostrzegany jako jeden z krytycznych czynników determinujących interakcję między miejscem II HSA, a niektórymi typami ligandów – małymi, zwykle aromatycznymi kwasami karboksylowymi.



Na podstawie wykresu do budowy modelu wybieramy dwie główne składowe.

- Równanie modelu

$$0.4406 + 0.0509 \cdot \log KCTAB + (-0.5246) \cdot CATS3D_00_DD$$

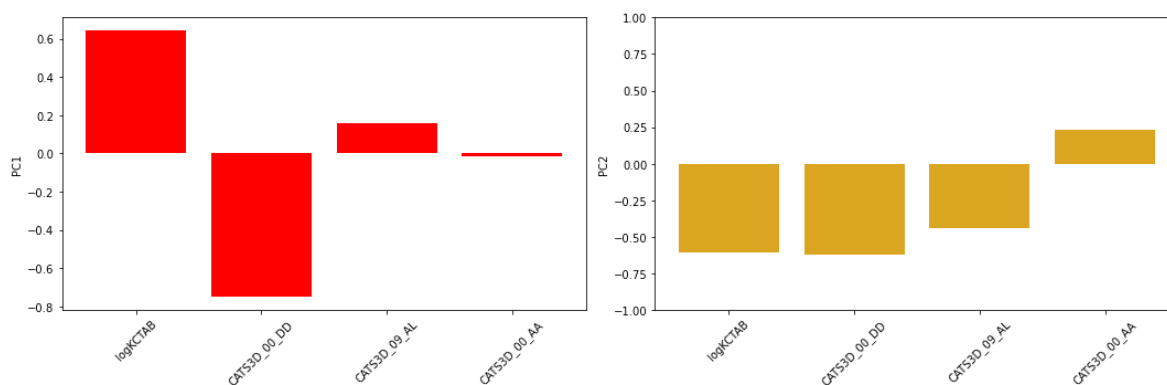
R^2 (współczynnik determinacji) równy 0.847 świadczy o dobrym dopasowaniu modelu do danych (im większy tym prosta regresji jest lepiej dopasowana do danych).

$Q2_{EX}$ (współczynnik walidacji zewnętrznej) równy 0.732 świadczy o dobrym modelu (im wyższa wartość (bliższa jedności) tym lepszy model.)

$RMSE_c$ (średni kwadratowy błąd kalibracji, miara jakości dopasowania modelu.) wynosi 0.448 – im mniejsza wartość tym model lepiej dopasowany.

$RMSE_{EX}$ (średni kwadratowy błąd przewidywania) wynosi 0.439 - lepszy model będzie miał mniejszy współczynnik $RMSE_{EX}$.

$RMSE_c$ oraz $RMSE_{EX}$ nie różnią się zbytnio, podobieństwo związków należących do zbioru kalibracyjnego nie jest zbyt duże. R^2 oraz $Q2_{EX}$ są bliskie jedności oraz nie różnią się pomiędzy sobą o więcej niż 0.3, co także dobrze świadczy o modelu.



Co prawda żaden z ładunków nie spełnia kryterium Malinowskiego – minimalna wartość (0.7) po pokonaniu której zachodzi korelacja pomiędzy zmienną ukrytą, a zmienną oryginalną, można jednak wyciągnąć następujące wnioski:

Pierwszy wektor własny (PC1) ma wysokie ładunki dodatnie przy zmiennej „logKCTAB” oraz wysoki ujemny ładunek przy zmiennej „CATS3D_00_DD”. Zmienne te są ujemnie skorelowane ze sobą.

Drugi wektor własny (PC2) wyjaśnia mniejszą część zmienności, a najwyższe ujemne ładunki występują przy zmiennych logKCTAB oraz CATS3D_00_DD. Zmienne te są dodatnio skorelowane ze sobą.

